Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime

(QED = Quality- and Efficiency-Driven)

Galit Bracha Yom-Tov

Queues in Hospitals: Stochastic Networks with ReEntering Customers in the QED Regime

(QED = Quality- and Efficiency-Driven)

Research Thesis

In Partial Fulfillment of the Requirements for Degree of Doctor of Philosophy

Galit Bracha Yom-Tov

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

TAMUZ, 5770 HAIFA JUNE 2010

The Research Thesis Was Done Under The Supervision of Prof. Avishai Mandelbaum in the Faculty of Industrial Engineering and Management.

I want to express my deep thanks to Prof. Avishai Mandelbaum for his professional guidance and trust in me.

"The best teacher is the one who suggests rather than dogmatizes, and inspires his listener with the wish to teach himself."

Edward Bulwer-Lytton (1805-1873)

The generous financial help of the Dan and Susan Rothenberg Doctoral Fellowship, the George and Lilian Schiff Technion Fellowship, the Forchheimer foundation fellowship, and the Israel National Institute for Health Policy and Health Service Research (NIHP) Doctoral Fellowship, the Technion is gratefully acknowledged.

To my loving family -
Zipora, Micha, and Elad.

Contents

List of Figures

List of Tables

Abstract

We study queues in healthcare. We start by developing and analyzing a queueing model, which we call Erlang-R, where the "R" stands for ReEntrant customers. The Erlang-R model accommodates customers who return to service several times during their sojourn within the system. It is most significant in time-varying environments. Indeed, it was motivated by healthcare systems, in which workloads are time-inhomogeneous and patients often go through a discontinuous service process. For example, in Emergency Wards, physicians are revisited by patients whose service process consists of cycles: examination by a physician, lab tests, treatment by a physician and so forth.

This thesis consists of three parts: open Erlang-R, semi-open Erlang-R, and Empirical analysis. In the first part, the main question we address is: how many servers (doctors/nurses) are required (staffing) in order to achieve predetermined service levels stably over time. Based on our theory, we propose a staffing policy that attains pre-specified service levels in the Halfin-Whitt (QED) regime. This policy applies the Modified Offered Load (MOL) approximation. We validate our policy, via simulation, both for large and small systems, and we use an EW simulator to validate its usefulness in realistic scenarios. We thus show how to stabilize, via proper staffing, both service levels and servers' utilizations, in time-varying healthcare environments.

In the second part, we concentrate on analyzing semi-open queuing networks with ReEntrant customers. These networks are used to model a Medical Unit with s nurses that cater to n beds, which are partly/fully occupied by patients. Here the questions we addressed here are: How many servers (nurses) are required (staffing), and how many fixed resources (beds) are needed (allocation) in order to minimize costs while sustaining a certain service level? We answer this by developing QED regime policies that are asymptotically optimal at the limit, as the number of patients entering the system (λ), the number of beds (n) and the number of servers (s) grows jointly. Our steady-state approximations turn out accurate for parameter values that are realistic in a hospital setting. We then use these approximations to develop MOL approximation to the closed-version of the Erlang-R model in a time-varying environment.

Our research was done in collaboration with one of the largest hospitals in Israel. This partnership provided us with the opportunity to analyze real data of patient-flow throughout the hospital, and validate our research in realistic situations. The last part of the research consists of this data analysis, concentrating mainly on hospitalization data in internal wards.

List of Abbreviations and Notation

Abbreviations

QED	Quality- and Efficiency-Driven
EW	Emergency Ward
MOL	Modified Offered Load
LOS	Length of Stay
MU	Medical Unit
IT	Information Technology
IW	Internal Ward
FTE	Full Time Equivalent
RN	Registered Nurse
AHA	American Hospital Association
M/M/s	(Erlang-C) a birth-death queueing model with infinite-capacity queue, Poisson arrivals, Exponential service times, and s servers
M/M/s/s	(Erlang-B) a birth-death queueing model with no queue, Poisson arrivals, Exponential service times, and s servers
M/M/s+M	(Erlang-A) a birth-death queueing model with infinite-capacity queue, Poisson arrivals, Exponential service times, s servers, and Exponential patience
FCFS	First Come First Served
i.i.d.	independent and identically distributed
OL	Offered Load
SIPP	Stationary Independent Period by Period
ISA	Infinite Server Approximation
QoS	Quality of Service
PSA	Piecewise Stationary Analysis
RCCP	Rough Cut Capacity Planning
ODE	Ordinary Differential Equation
QD	Quality Driven
ED	Efficiency Driven
LWBS	Left Without Being Seen
NRP	Nurse Rostering Problem
NSP	Nurse Scheduling Problem

IVR	Interactive Voice Response
ALOS	Average Length of Stay
WIP	Number of Patients in Ward
V	Ventilated
ICU	Intensive Care Unit
DOW	Day Of Week
FSLLN	Functional Strong Law of Large Numbers
FCLT	Functional Central Limit Theorem
DE	Differential Equation
CLT	Central Limit Theorem
u.o.c.	uniformly on compact
a.s.	almost surely
BM	Brownian Motion
HRM	Human Resources Management

Notation

Part I - Notation

$s(t)$	Number of doctors at time t
$\lambda(t)$	Arrival rate at time t
p	Probability of staying in the medical unit after service
G_1	General distribution function of *Needy* state (station 1)
μ	Service rate (in *Needy* state (1))
G_2	General distribution function of *Content* state (station 2)
δ	Content/delay rate
S_i	Service/delay time in station i
$R_i(t)$	Offered Load in station i at time t
S_i^{*j}	Sum of j independent random variables S_i
β	QED quality of service parameter
$Q_i(t)$	Number of patients in station i at time t
E	Expectation
P	Probability measure

P_{ij}	Stationary probability that there are i patients in station 1 and j patients in station 2
R_i	Steady-state offered load of station i
ρ	Offered load per server; $\rho = \frac{R_1}{s} = \frac{\lambda}{(1-p)s\mu}$
$P(W > 0) = \alpha$	Probability of waiting
$P(W > t)$	Probability of waiting more than t
$E[W]$	Expected waiting
$\phi(\cdot)$	The standard normal density function
$\Phi(\cdot)$	The standard normal distribution function
$S_{i,e}$	Excess service time at station i
$\lambda_i^+(t)$	Aggregated arrival rate function to node i at time t
VAR	Variation
$\bar{\lambda}$	Average arrival rate
κ	Relative amplitude of Sinusoidal arrival rate function
f	Period of Sinusoidal arrival rate function
ω	Frequency of Sinusoidal arrival rate function
$Amp(\cdot)$	Amplitude of periodic function $\cdot$
$Phase(\cdot)$	Phase shift of periodic function $\cdot$ with respect to the entrance arrival rate
$R^C(t)$	Offered Load of $M_t/M/s_t$ (time-varying Erlang-C) model at time t
$Q_i^{(0)}(t)$	Fluid solution of the number of patients in station i process, at time t

Part II - Notation

s	Number of nurses
n	Number of beds
λ	Arrival rate
μ	Service rate
δ	Dormant/activation rate
γ	Cleaning rate
p	Probability of staying in the medical unit after service
β	First QED quality of service parameter
η	Second QED quality of service parameter
r	Jennings and de Vericourt parameters ratio
$N(t)$	The number of *needy* patients at time t

$D(t)$	The number of *dormant* patients at time t
$C(t)$	The number of beds in *cleaning* at time t
$\pi(i,j,k)$	The stationary probability of having i *needy* patients, j *dormant* patient and k beds in *cleaning* (sometimes denoted $\pi_n(i,j,k)$ or $\pi_{n,s}(i,j,k)$)
P_l	The probability that there are l beds occupied in the system
$\pi^A(x-e_i)$	The probability that the system is in state $x-e_i$ at the **arrival epoch** of a customer to node i
$P(W=0)$	Probability to get immediate service
$P(W>0)$	Probability of waiting/delay
$P(blocked)$	Probability of blocking of the medical unit ($P(blocked)=P_n$)
$E[W]$	Expected waiting
W	Steady-state in-queue waiting time, for a hypothetical newly needy patient
$p_n(s,t)$	Tail of the steady state distribution of W
$OC(n,s)$	Average occupancy level
ρ	Offered load per server; $\rho=\frac{\lambda}{(1-p)s\mu}$
R_N	The solution of the balance equations for the *needy* state; $R_N=\frac{\lambda}{(1-p)\mu}$
R_D	The solution of the balance equations for the *dormant* state; $R_D=\frac{p\lambda}{(1-p)\delta}$
R_C	The solution of the balance equations for the *cleaning* state; $R_C=\frac{\lambda}{\gamma}$
A	Sum of non-service stations offered load; $A=R_C+R_D$
B	Offered load ratio; in IW model :$B=\frac{R_N}{R_C+R_D}$, in semi-open Erlang-R model : $B=\frac{R_1}{R_2}$
$\approx$	$a_n \approx b_n$ if $a_n/b_n \to 1$, as $n \to \infty$
$\phi(\cdot)$	The standard normal density function
$\Phi(\cdot)$	The standard normal distribution function
$N(0,1)$	A standard normal random variable with distribution function Φ
E	Expectation
P	Probability measure
C_n	Annual bed costs
C_s	Annual nurse costs
$s(t)$	Number of nurses at time t
$R_i(t)$	Offered Load in station i at time t

Part II - Notation

T_i	Average time that the system is in state i
$P_{i,j}$	Probability to transfer from state i to state j, given that one is presently in state i
h_t	Probability to leave the medical ward in day t given that LOS is at least t

1 Introduction

1.1 The Structure of Modern Hospitals

A Hospital or Medical Center is an institution for health care, which is able to provide long-term patient stays. One distinguishes between two types of patients: inpatients and outpatients. Some patients in a hospital come only for a diagnosis and/or therapy and then leave ('outpatients'), while others are 'admitted' and stay overnight or for several weeks or months ('inpatients'). Hospitals usually differ from other types of medical facilities by their ability to admit and care for inpatients. Within hospitals, the two types of patients are usually treated in separate systems, and thus can be analyzed separately. We will concentrate on the inpatient system.

In the modern age, a hospital constitutes a combination of several Medical Units (MU) specializing in different areas of medicine such as internal medicine, surgery, plastic surgery, and childbirth. In addition to these medical units, the hospital includes some service units such as laboratories, imaging facilities, and IT (Information Technology), that provide service to the medical units. Typically, inpatients arrive to the hospital, randomly, via an Emergency Ward (EW), which deals with immediate threats to health and has the capacity to exercise emergency medical services. For operational purposes, therefore, the flow of patients in a hospital can be viewed as in Figure 1: a patient enters the EW, is treated, and then discharged after treatment or admitted to stay; the latter if the doctors decided to hospitalize the patient and there is an available bed at an appropriate MU, in which case the patient is transferred to that MU. At some point in time (i.e., when the patient is cured or transfered to other medical centers, or unfortunately dies) the patient leaves the hospital system.

Focusing on the operational point of view, the hospital includes doctors, nurses and administrative staff. Each MU is managed autonomously, with its own medical staff. Each MU has a limited capacity which is a function of the physical space (static capacity) and the staffing levels (dynamic capacity). The physical space is usually measured by the number of beds allocated to that MU, and the staffing levels by the number of service providers: doctors, nurses, and general workers (sanitation staff, etc.). Naturally, capacity restrictions can lead to a situation of system blocking. Thus the EW and MUs can be blocked, and a situation where ambulances are turned away [25], or a patient is waiting in the EW for assignment is not a rare event. In large medical centers there are several MUs of the same type. This division is due to a combination of location constraints and the inability to manage large wards efficiently. Nevertheless, blocking does also occur in such large medical centers.

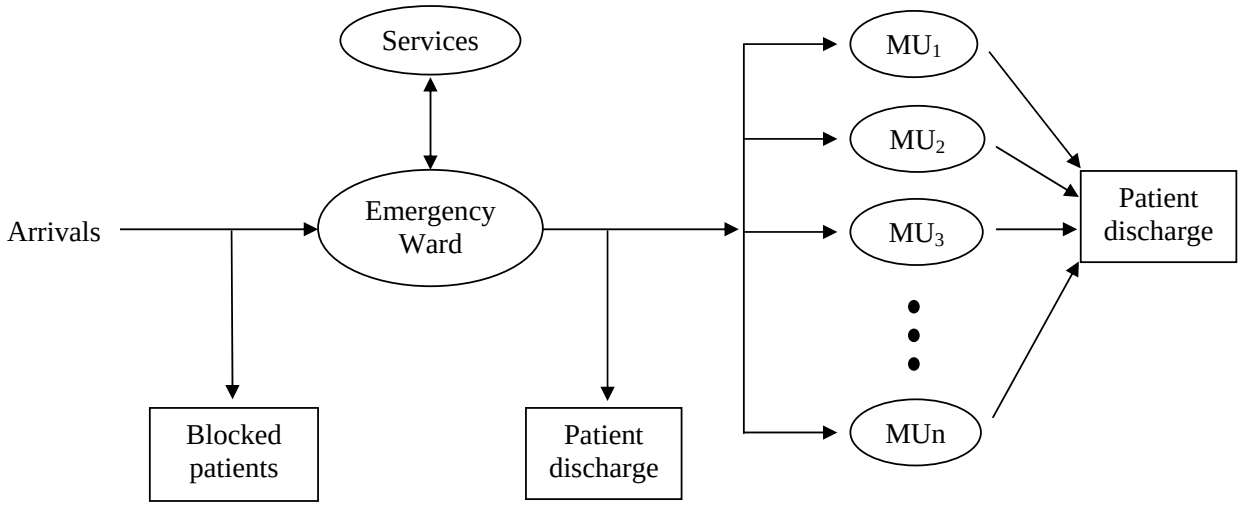


Figure 1: The basic operational model of a hospital system

All of the above leads to the conclusion that one can model the hospital as a complex stochastic network, where each node represents some process. We can then examine the flow of patients in that network, as shown in the case-study of de Bruin et al. [19] and Hall [41]. If that flow is not smooth then patients are delayed at various points in the system, waiting for medical care or waiting in queues for other reasons. The most notorious queues for medical care are those for surgery, organ transplants, very expensive diagnostic tests such as C.T. and MRI, and specialists. For some of the above, the wait can be as long as several months [15]. Less noticed, though much more common, are queues for hospital beds, doctors, nurses, lab tests and dispensing medication. These queues cause delays during treatment, when the response time can be critical for patient safety and quality of care (see Sobelov et al. [69]).

Healthcare queueing research has the capability to deal with various aspects of the healthcare system. For example:

1. Scheduling - for example, the optimal scheduling of surgery rooms, in order to minimize wait while considering diverse patients' needs and system constraints; or managing out-patient appointments [41].

2. Routing - for example, the routing of patients from the EW to the MUs [71].

3. Staffing - for example, how many nurses to assign to an MU [45], first at the planing stage and then dynamically.

4. Design - for example, capacity planning [32] and what is the optimal bed allocation [31].

5. Costing - for example, the optimal sharing of surgical costs in the presence of queues [29].

Some of these issues have been noticed and approached in the past, usually not as a healthcare problem but rather from a more general perspective. Many aspects, however, have not been treated, and some of them are crucial for healthcare systems, such as adaptivity of large-system-approximations to small systems, and the combination of medical and psychological aspects. The most common method for modeling healthcare systems is simulation (see for example [58]). The reasons for the popularity of simulation in healthcare seem to the same as that of call centers: there is a widening gap between the complexity of the modern healthcare system and the analytical models available to accommodate this complexity. Moreover, simulation techniques are relatively simple user-friendly tools [28]. We aim to narrow this gap by developing simple queueing models that capture the ReEntrant effect of patients, and show how these relatively simple models can actually capture enough of the system dynamics, and therefore can be used to model complex EW and MUs environments. There are other researchers who use various methods of mathematics for modeling and analysis of healthcare systems (see Halls' book [41] for works on the subject). Only a minority of them have tried to deal with the stochastic characteristics of the system, using queueing theory, similarly to outpatient analysis and in other fields such as call centers [28]. Nevertheless, even this small body of work suggests that stochastic-models insights could significantly advance our understanding of inpatient healthcare systems.

Healthcare systems are highly regulated, both in service and operational issues. Each country provides the regulator with different means to control healthcare systems: for example, Israel has a national health law. The Ministry of Health controls all beds allocation and staffing capacity in hospitals. In Europe, some countries regulate maximum queues waits for surgeries, and in 2004 in the US, the California Department of Health Services (CDHS) published a law that specifies nurse-to-patient ratios that determine the minimal staffing levels allowed [63]. Some of these laws are poorly designed, as Jennings and de Véricourt [45] showed concerning the California regulations. Queueing models can be used to better design staffing regulation, as shown by Green et al. [37]. We believe that our work will provide the regulator tools to better understand the impact of capacity-related decision on service level in hospitals. This will help change the current practice of determining staffing and allocation based on utilization and budget constrains, to a more balanced approach that accounts for service aspects as well. There are still some research questions to be answered before one is able to actually establish staffing guidelines that are based on our research; for example, healthcare systems are time-varying and it is unclear over what period (year, week, or day) one

should base such guidelines.

1.2 Research Objectives and Contributions

We concentrate on capacity management problems in hospitals for several reasons: First, the main resource used in healthcare systems, as in many other service industries, is the human resource. Doctors, nurses, therapists, laboratory technicians, and so forth are the main resources of that system and their salaries constitute 70% of hospital expenditure [62]. The second reason is that these personnel have very long training periods, and healthcare systems suffer from a chronical shortage of medical personnel, which has a detrimental impact on the system. For example, in the US alone, in 2005 there were 1.1 million FTE (Full Time Equivalent) Registered Nurse (RN) jobs [1], but there is still a chronic shortage of nurses; the American Hospital Association (AHA) reported that US hospitals had an estimated 116,000 RN vacancies as of December 2006 [2], and that the personnel shortage causes some very serious problems in the majority of hospitals: decreased staff satisfaction (in 49% of the hospitals), EW overcrowding (36%), diverted EW patients (35%), reduced number of staffed beds (17%), increased waiting times to surgery (13%), and more.

Hence, we seek to support capacity management policies for healthcare systems, using stochastic processes. We develop strategies for doctor-staffing in the time-varying environment of Emergency Wards (see Part I), and a joint nurse-staffing and bed-allocation strategy for Medical Wards (see Part II). Our service-level objectives reflect both blocking phenomena and the response time of the medical staff. We develop QED (Quality and Efficiency Driven) staffing policies that balance between these service-level objectives and the efficiency of the system. We validate our models using simulation in a realistic setting. We also include, in Part III of this work, empirical analysis of patients flow data. This analysis describes some of the statistical characteristics of the flow of inpatients through the MUs, such as the arrival rates, LOS, blocking etc.

The models we describe in this work have limitations, which call for natural extensions. Examples of such extensions include multi-classes of patients (to distinguish service phases or medical priorities), abandonments (customers who leave without being seen), and random parameters. We elaborate on some of these extensions in Part IV.

The main contribution of this work is the addition of a new central feature to the queueing literature, which incorporates returning customers to service. We called our model Erlang-R, were "R" stands for Returning or ReEntrant. Within the Erlang-R framework (open, semi-open), we identified the circumstances where returns must be modeled explicitly, as opposed to being "absorbed" in the corresponding simpler setups (M/M/s, M/M/s/n). Regarding the above-mentioned

circumstances, we have contributions along three main directions:

1. The QED regime: Generalizing the QED limits, and exact calculations of Khudyakov et al. [50] to semi-open service system.

2. Coping with time-varying environments:

   - Developing MOL (Modified Offered Load) approximations, and analyzing insightful special-cases.
   - Stabilizing performance using the MOL/ISA approach of Feldman et al. [26]:
     - Developing generalization of the MOL-QED staffing procedure for open and semi-open queueing networks.
     - Validating the usefulness of this procedure in small scale systems.
     - Using simple models to stabilizing complex (real) systems.
   - Fluid and Diffusion approximations:
     - Developing time-varying fluid and diffusion approximations within the framework of Mandelbaum et al. [53] .
     - Develop corrections to those processes under MOL-QED staffing.

3. Open vs. Semi-open networks:

   - Identifying the significance and understanding the joint combining impact of time-varying arrivals parameters (amplitude and phase), and service-rate on open-system performance measures.
   - Identifying the significance and understanding the impact of the offered-load ratio, on semi-open system performance measures.

Part I

The Open Erlang-R Model

2 Introduction: The Erlang-R Model

It is natural to use queueing models to support workforce management in service systems. Most common are the Erlang-C (M/M/s), Erlang-B (M/M/s/s) and Erlang-A (M/M/s + M) models, all used in call centers. But when considering healthcare environments, we find that these models lack a central prevalent feature, namely, that customers might return to service several times during their sojourn within the system. Therefore, the service offered has a discontinuous nature and is not provided at one time. This has motivated our queueing model, Erlang-R ("R" for ReEntrant customers) which accommodates the return-to-service phenomena.

More explicitly, we consider a model where customers seek service from servers. After service is completed, with probability $1-p$ they exit the system and with probability p they return for further service after a random delay time. We refer to the service phase as a *Needy* state, and to the delay phase as a *Content* state (following Jennings et al. [46]). Thus, during their stay in the system, customers start in a Needy state and then alternate between Needy and Content states. We assume that there are s servers in the system. When customers become Needy and an idle server is available, they are immediately treated by a server. Otherwise, customers wait in queue for an available server. The queueing policy is FCFS (First Come First Served).

We assume that the Needy service times are independent and identically distributed (i.i.d.), with general distribution G_1 and mean $\frac{1}{\mu}$, and that the Content times are also i.i.d. with general distribution G_2 and mean $\frac{1}{\delta}$. We also assume that the Needy and Content times are independent of each other and of the arrival process. The arrival process is taken to be a time-inhomogeneous Poisson process with rate $\lambda_t,\ t \geq 0$. Some of our results require that the Needy and Content times are exponentially distributed. We shall state specifically when this is the case.

Figure 2 displays our system graphically:

Note that our returning customers are different from the redialing customers of a call center. The latter leave the system before service in response to a busy line or due to abandonment. See Aksin et al. [3] and Artalejo et al. [6] for further details of such models.

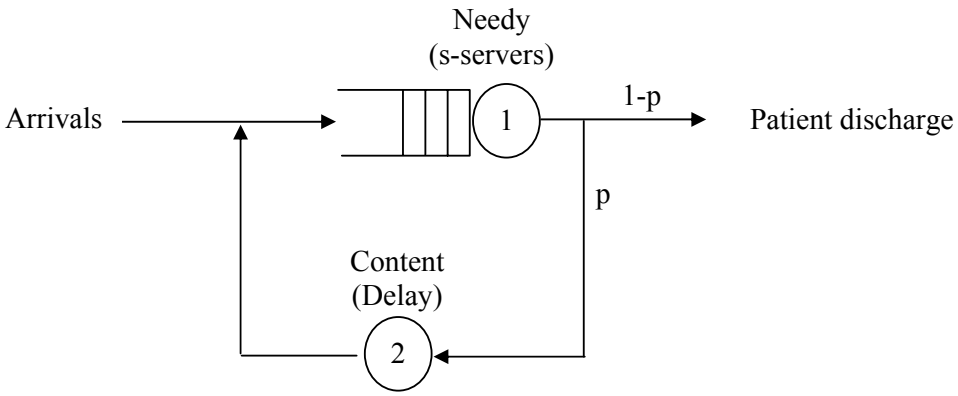


Figure 2: The Erlang-R model

2.1 Examples in Healthcare

We now describe a few examples where the Erlang-R model is applicable in hospitals. The first example presents the process of doctor service (or nurse service) in an EW. Patients enter the EW, and are referred to a doctor. The doctor examines them, and decides whether to send them home or to admit them to the hospital. In most cases, the decision is made after the patient goes through a series of medical tests. Thus, the process that a patient goes through, from the doctor's perspective, fits our model. A patient visiting the doctor is in a Needy state. Between each visit, the patient is considered to be in a Content state, which represents the delay caused undergoing medical tests such as X-rays, blood tests, and examinations by specialist. After each visit to the doctor, a decision is made to release the patient from the EW (either to his/her home or to the hospital), or to direct the patient to additional tests.

A second example is the Radiology reviewing process [51]. After a mammography test, the radiologist interprets the results. This includes several stages: examining referral requisition, reviewing clinical background information, analyzing images, and dictating results. In some cases, part of the information on the patient is lacking: the radiologist does examine the case but it must be put on hold, waiting for this additional information to arrive; after arrival, the reviewing process starts again. With radiologists being the servers, this can be modeled using our Needy-Content cycle.

The final example is the process of bed management in an Oncology Ward. In such a medical ward, patients return for hospitalization and treatment, much more frequently than in regular wards. Here servers are the beds, the Needy state models the times when a patient is in the hospital, and the Content state models the times when the patient is at home. A patient leaves the system when cured or unfortunately passes away.

We would like to understand the significance of customers' service cycles. The fact that service is not given in one time but, rather, separated into several visits to the server, could affect operational decisions: we seek to understand in which cases it does, and what are the implications regarding staffing procedures.

2.2 Main Results

In this part we first show that, in *steady-state*, our model behaves like an Erlang-C (M/M/s) model. This applies from the quality-of-service perspectives in the Needy state, since the marginal distribution of the number of customers in the Needy state is the same as in steady-state Erlang-C. Nevertheless, the Erlang-R model is useful beyond Erlang-C for two reasons: first, it provides more information than Erlang-C since we model the Content state as well. But more importantly, we found that while steady-state performance is identical, there is a significant difference between the two models as far as transient behavior is concerned; in particular, in transient times, the Content state plays an important role in determining appropriate staffing levels. This finding is important since the systems in our healthcare examples, as well as in many other systems, are typically in a transient state due to the nature of the arrival process, the rate of which varies significantly in time.

Staffing systems that are in a transient state differ from those in steady state. Instead of setting the service-quality measures in the long run, one must consider them at every moment in time. Our goal is to identify staffing procedures that stabilize performance over time. Specifically, no matter what time of day customers enter the system, they will always wait on average for the same amount of time, and their probability of waiting remains constant. Thus, the staffing algorithm is to attain pre-specified service levels but, at the same time, servers' utilization must be high. This means that we seek to create QED (QED = Quality and Efficiency Driven) balance, at all times of a transient system.

We use the MOL (Modified Offered Load) approach [44] in which one first calculates the time-varying *offered load*. In our case, the offered load for Station 1 (Needy) and Station 2 (Content) are given (Section 5) by

$$R_1(t) = E\left[\sum_{j=0}^{\infty} p^j \lambda(t - S_1^{*j} - S_2^{*j} - S_{1,e})\right] E[S_1],$$
$$R_2(t) = E\left[\sum_{j=1}^{\infty} p^j \lambda(t - S_1^{*j} - S_2^{*j-1} - S_{2,e})\right] E[S_2].$$

Here S_i^{*j} is the sum of j independent random variables S_i, where S_1 is the Needy service time and S_2 is the Content time; their joint distribution is given by the convolution of their separate distributions. These formulae are hard to calculate. We hence develop ways to approximate and estimate $R_1(t)$ and $R_2(t)$ and calculate them numerically (for example by using Taylor-series approximation) - see Section 5. In some cases, a closed form solution is available. This enables one to treat some cases analytically, which gives rise to managerial insights on our system.

In the MOL QED approach [26], staffing is determined by substituting the time-varying offered load formula into the square-root staffing formula:

$$s(t) = R_1(t) + \beta\sqrt{R_1(t)},\ t \geq 0.$$

We demonstrate that this approach works very well. We find that, in most cases, performance measures such as the probability of timely service, expected waiting, and servers' utilization are all remarkably stable over time. The reason for success is that time-varying square-root staffing controls the system, at all times, in a state that is very close to a naturally-corresponding steady state system. This also explains why the constant β is calculated using steady-state formulae, and it does not vary in time. We show that, although our staffing algorithm is based on large-scale approximations, it also stabilizes small systems such as those in hospitals, where the number of 'servers' (e.g. doctors) varies between 1 and 10.

We demonstrate the importance of using staffing based on the time-varying Erlang-R model (Section 6). In one stylized example, the arrival rate function is given by

$$\lambda_t = 30 + 30 * 0.2 * \sin\left(\frac{2\pi}{24}t\right),\ t \geq 0,$$

with the parameters $p = 2/3$, $\mu = 1$, $\delta = 0.5$. Introducing the following service goal: $P(W_t > 0) = 0.5$, Figure 3 presents the probability of waiting when one uses the Erlang-R, Erlang-C and PSA algorithms for staffing. It clearly shows that, while using Erlang-R stabilizes system's performance around the pre-specified target, using Erlang-C or PSA does not. Other realistic healthcare examples can be found in Sections 6.2 and 7.

Investigating the differences between Erlang-C and Erlang-R revealed the environments in which Erlang-R is essential. We have already mentioned the main characteristic that is time-varying arrivals. More explicitly, we show that the difference manifests itself both in amplitude and phase of the offered load which, in turn, is the driver of the system's dynamics. For example, when the arrival rate is periodic, and the service times are exponentially distributed, the amplitude of an Erlang-R Offered Load (OL) is always smaller than the amplitude of a corresponding Erlang-C OL.

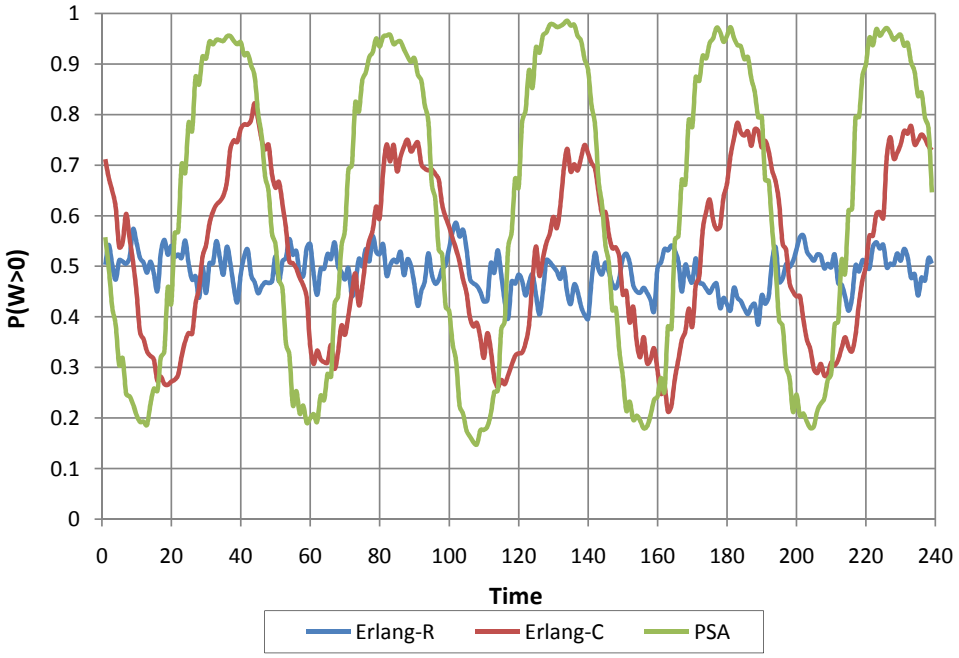


Figure 3: $P(W > 0)$, as a function of time, when staffing according to Erlang-R, Erlang-C, and PSA

On the other hand, the phase of Erlang-R OL sometimes leads and sometimes lags behind the phase of Erlang-C OL. In fact, these differences between the two models are especially pronounced when arrivals vary during the sojourn time of a customer, which is exactly the case in emergency wards.

The implication of miscalculating the offered load is that Erlang-C will lead to over- or under-staffing at most times. One must thus take into account the discontinuous nature of service, in order to avoid excessive staffing costs or undesirable service levels. Using Taylor-series approximations, we can quantify the differences between the two models also for general arrival-rates functions and general service-time distributions.

Lastly, based on diffusion approximations, we developed new MOL approximations for the number of Needy customers and the expected waiting time in the QED regime, which are also highly accurate.

The subsequent sections are organized as follows: first we review some of the related literature; then, in Section 4, we describe the steady-state of the Erlang-R model, in particular, showing that its Needy part behaves exactly as Erlang-C (M/M/s). In Section 5 we introduce the time-varying offered-load, which will be used later in time-varying staffing procedures for our model. In that section, we present a numerical method for calculating the offered load when service times are exponentially distributed, as well as approximations using Taylor expansions, for general service

time distributions. In addition, we analyze the offered load for periodic arrival rates (sinusoidal), and compare the Erlang-R and Erlang-C models in order to determine the circumstances in which Erlang-R is mostly needed. In Section 6 we validate our time-varying staffing procedure, using simulation, and show that it achieves pre-specified service-level requirements. In Section 7 we present an application of Erlang-R in an EW environment. Lastly, in Section 8 we develop MOL approximations for the number of customers in each state, and the average waiting time process in the QED regime.

3 Literature Review

3.1 Staffing Problems in Hospitals

The workforce of a hospital consists of nurses, doctors, laboratory workers and others. Most of these human resources require long and costly training, and jointly contribute as much as 70% to a hospital's operational budget [62]. Nurses' salaries make up the largest single element in hospital costs [67]. Thus, careful management of work-force capacity is naturally called for.

Queueing models help determine personnel levels that should be available to serve patients over a given time slot. These staffing levels must vary during a day as they track predictable variations in the arrival rates of patients. The prevailing schemes, however, are inflexible; for example, the application of beds-to-personnel ratios is common when considering nurses staffing [63]. There do exist queueing models for staffing personnel, but most account neither for time-varying environments nor recurrent services. The first to consider the effect of returning patients in healthcare were Jennings and de Véricourt [45]. They used a closed queueing model to develop new recommendations for nurse-to-patient ratios which are a scale-dependent, being developed in the QED regime. Yom-Tov and Mandelbaum [73] then expanded [45] to accommodate bed allocations. But both [73] and [45] impose a restriction on the number of patients (which is not the case here), and they analyzed the system in steady-state. To the best of our knowledge, the only exceptions to consider explicitly time-varying queues in hospitals are those of Green et al. [34, 35], Bekker and de Bruin [10], and Zeltyn et al. [74]. Green et al. apply the Erlang-C model, and the Lag-SIPP (Stationary Independent Period by Period) approach for staffing doctors in the EW, with the arrival rate varying during the day. Our goal is to demonstrate that Erlang-R is more appropriate for modeling the time-varying EW environment, which is due to the discontinuous nature of service during patients' stay. Bekker and de Bruin use loss systems for analyzing time-varying weekly patterns on beds' allocation. Zeltyn et al. apply the ISA (Infinite Server Approximation) method of [26] (see below), plus heuristics, to determine physicians' staffing. Finally, we refer the reader to Green et al. [34] comprehensive survey of time-varying queues and their applications in workforce management.

3.2 The QED (Quality- and Efficiency-Driven) Regime

We shall focus on QED queues in order to balance patients clinical needs for timely service with the economical preferences to operate at high efficiency. This operating regime is characterized by high levels of resource-utilization *jointly* with high service-quality. The latter is characterized by short queueing delays, being one order of magnitude shorter than service durations or, equivalently, by a

significant fraction of customers (e.g. 30 – 70%) who get served immediately upon arrival.

QED models have become popular for research as they capture the operational environment of call centers [28]. However, as first noticed by Jennings and de Véricourt [45] QED approximations are already useful for the much smaller scale healthcare environment, a fact that was recently substantiated formally in [43].

QED queues adhere to some version of the so-called *square-root staffing rule*. For example, QED staffing in an Erlang-C (M/M/s) model corresponds to the number of servers s being $s \approx R+\beta\sqrt{R}$; here R is the *offered load*, given by $R = \lambda \cdot \mathrm{E}[S]$ (λ is the arrival rate and $\mathrm{E}[S]$ is the mean service time), and β is a QoS (Quality-of-Service) parameter that is set to accommodate service-level constraints. The square-root staffing rule was described by Erlang [23], as early as 1924. However, its formal analysis awaited the seminal paper by Halfin and Whitt [39], in 1981.

3.3 Staffing Time-Varying Queues

When the arrival rate varies with time, the above QED approach requires modifications. As mentioned already, we aim the staffing algorithm to attain a pre-specified service level *stably over time*, while maintaining a high level of servers' utilization, as is typically the case in the QED regime.

This goal has been addressed via two approaches. The first tries to find the right staffing level using steady-state approximations, such as in PSA (Piecewise Stationary Analysis), RCCP (Rough Cut Capacity Planning) [74], SIPP, or lag-SIPP [44, 33, 35]. In PSA, for example, we divide the time-horizon into small planning intervals and calculate the average arrival rate for each interval. Then, assuming that the system gets fast to steady state, and using that average arrival rate, we calculate the steady-state offered-load for each interval (i.e., $R(t) \approx \bar{\lambda}(t)E[S]$), then staff according to steady-state recommendations using square-root staffing.

The second approach includes algorithms such as MOL (Modified Offered Load) [44] or ISA (Infinite Server Approximation) [26]. Here, one tries to calculate the time-varying offered load, using a corresponding system with ample servers. For example, in the time-varying Erlang-C model $(M_t/M/s_t)$, when using the MOL approach, we calculate the time-varying offered load [22] via

$$R(t) = E\left[\lambda(t-S_e)\right]E[S] = E\left[\int_{t-S}^{t}\lambda(u)du\right] = \int_{-\infty}^{t}\lambda(u)P(S>t-u)du.$$

(This is the number of customers in a corresponding $M_t/G/\infty$ queue.) Then we use an adaptation of the square-root formula: $s(t) = R(t) + \beta\sqrt{R(t)}$. This approach works very well, as shown in Jennings et al. [44] and Feldman et al. [26], and we are following it here with our Erlang-R model. Research is still needed to provide theoretical justification for why the MOL approach actually

works. The only available theoretical support is Feldman et al. [26] who proved convergence of the diffusion process under MOL staffing, for the Erlang-A model with patience rate equals to service rate. An analysis of the time-varying offered load in the Erlang-C environment was carried out by Eick et al. [21, 22]. They also developed Taylor-series approximations for cases in which the offered load cannot be calculated explicitly. We use their work in Section 5.3, to compare Erlang-C with Erlang-R.

4 Steady-State Performance Measures

We start with a steady-state analysis of the Erlang-R model. This entails simple calculations for a two-state Jackson network, which provide the backbone for later analysis. In particular, it comes out that, in steady state, the probability of waiting depends exclusively on the offered load of the Needy station. In addition, this probability has exactly the same structure as in a standard Erlang-C (M/M/s) model. We provide formulae for all the standard quality measures. In addition, our model provides more information about the system than does Erlang-C, since it accounts for the delayed customers in the Content state as well. One can use this information to approximate the number of customers in the system, as we will discuss later.

In this section, we assume that the service times are exponentially distributed with parameters μ and δ, and that the arrival rate is constant $\lambda(t) \equiv \lambda$. Let $Q = \{Q(t), t \geq 0\}$ be a two-dimensional stochastic queueing process, where $Q(t) = (Q_1(t), Q_2(t))$: $Q_1(t)$ represents the number of *Needy* patients in the system, and $Q_2(t)$ the number of *Content* patients at time t. Under our assumptions, the system is an open (product-form) Jackson network with the following steady state distribution:

$$P_{ij} := P\left(Q_1(\infty) = i, Q_2(\infty) = j\right) = \frac{(R_1)^i}{\nu(i)}\pi_{01}\frac{(R_2)^j}{j!}\pi_{02},$$

where

$$\pi_{02} = \left[\sum_{j=0}^{\infty}\frac{(R_2)^j}{j!}\right]^{-1} = e^{-R_2},\ \pi_{01} = \left[\frac{(R_1)^s}{s!(1-R_1/s)} + \sum_{i=0}^{s-1}\frac{(R_1)^i}{i!}\right]^{-1},$$

$\nu(i)$ is defined as

$$\nu(i) := \begin{cases} i! & , \quad i \leq s, \\ s!s^{i-s} & , \quad i \geq s, \end{cases}$$

and

$$R_1 = \frac{\lambda}{(1-p)\mu},\ R_2 = \frac{p\lambda}{(1-p)\delta}.$$

We call R_1 and R_2 *the steady-state offered load* of Station 1 and 2, respectively.

Theorem 1. *When the Needy service times are exponentially distributed (μ), and the arrival rate is constant (λ), then:*

$$\alpha := P\left(W > 0\right) = \left[\frac{(R_1)^s}{s!(1-R_1/s)}\right]\pi_{01},$$

$$P\left(W > t\right) = \alpha e^{-s\mu(1-R_1/s)t},$$

$$E[W] = \frac{\alpha}{\mu s\left(1-R_1/s\right)}.tkubho$$

Proof: see Appendix A.1 on page 148.

Note that the above measures depend exclusively on the offered load of the Needy station. Also note that these quantities have exactly the same structure of the Erlang-C (M/M/s) model. The difference is between the offered load measure. Here $R_1 := \frac{\lambda}{(1-p)\mu}$, while in Erlang-C, $R = \frac{\lambda}{\mu}$. Recall that $\frac{1}{1-p}$ is the expected number of visits in the Needy station. Therefore, in steady state, Erlang-R behaves exactly as Erlang-C with service rate $(1-p)\mu$, i.e. as if services were concatenated to one another, with no delay between them; consequently, it will have exactly the same QED approximations. Thus, we can conclude that, in steady-state, the appropriate QED staffing policy for our model is to set s such that

$$s = \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}},\ \beta > 0,$$

where β is given by the Halfin-Whitt formula [39]; it is related to the desired α by $\alpha = \left[1 + \beta\frac{\Phi(\beta)}{\phi(-\beta)}\right]^{-1}$, where $\phi(\cdot)$, and $\Phi(\cdot)$ are the standard Normal density and distribution functions, respectively.

However, not all the performance measures depend exclusively on the Needy state. Some incorporate information both of the Needy and Content states. For such measures Erlang-R and Erlang-C yield different results. For example, one might wish to evaluate the probability of overcrowding events in the EW. To do that, one might assess the probability that the number of patients in the EW will exceed the number of beds. This includes Needy patient as well as Content ones. In more general terms, we are interested in the probability that there will be more than n customers in the system, when $n \geq s$. This could be calculated by the following formula:

$$P\left(Q_1(\infty) + Q_2(\infty) \geq n\right) = \sum_{i=0}^{\infty}\sum_{j=n-i}^{\infty} P_{ij} = \sum_{i=0}^{\infty}\frac{(R_1)^i}{s!s^{i-s}}\pi_{01}\sum_{j=n-i}^{\infty}\frac{(R_2)^j}{j!}e^{-R_2}.$$

5 The Offered Load

As mentioned earlier, research has shown that appropriate staffing levels in non-stationary systems can be based on the offered load (Feldman et al. [26]). Adopting this approach, we now introduce the offered load of our time-varying Erlang-R, $R = \{R(t), t \geq 0\}$. We set $R(t) = (R_1(t), R_2(t))$ where $R_i(t)$ corresponds to node i in the network. We define R using a related system, with the same structure, in which the number of servers in node 1 is infinite, which results in an $(M_t/G/\infty)^2$ network. The offered load in our system equals the mean number of busy servers (equivalently, the number of served customers) in each node of this network. Massey and Whitt [59] showed that, in such networks, the average number of customers at node i is given by:

$$R_i(t) \equiv E[Q_i^{\infty}(t)] = E\left[\int_{t-S_i}^{t} \lambda_i^+(u)du\right] = E[\lambda_i^+(t - S_{i,e})]E[S_i] \tag{5.1}$$

where λ_i^+ is the aggregated-arrival-rate function to node i, and $S_{i,e}$ is a random variable representing the excess service time at node i. Note that (5.1) is valid for general service time distributions, and that if S_i is exponentially distributed, then $S_{i,e} = S_i$.

In our two-node network, λ_i^+ is defined by the minimal non-negative solution to the system traffic equations

$$\lambda_1^+(t) = \lambda(t) + E[\lambda_2^+(t - S_2)],$$
$$\lambda_2^+(t) = pE[\lambda_1^+(t - S_1)].$$

These equations constitute a variation of Fredholm's integral equation [64], which one can solve recursively (using the fact that S_1 and S_2 are independent) as follows:

$$\lambda_1^+(t) = \lambda(t) + pE[E[\lambda_1^+(t - S_2 - S_1)]] = ... = \sum_{j=0}^{\infty} p^j E[\lambda(t - S_1^{*j} - S_2^{*j})],$$
$$\lambda_2^+(t) = pE[\lambda(t - S_1) + E[\lambda_2^+(t - S_1 - S_2)]] = ... = \sum_{j=1}^{\infty} p^j E[\lambda(t - S_1^{*j} - S_2^{*j-1})].$$

Here S_i^{*j} is the sum of j i.i.d random variables S_i, hence its distribution is the j-convolution of S_i. The above representation of $\lambda^+(t) = (\lambda_1^+(t), \lambda_2^+(t))$ reveals that it is an infinite sum of delayed arrival rates.

Substituting $\lambda^+(t)$ into $R(t)$ yields

$$\begin{aligned} R_1(t) &= E[\lambda_1^+(t - S_{1,e})]E[S_1] = E\left[\sum_{j=0}^{\infty} p^j \lambda(t - S_1^{*j} - S_2^{*j} - S_{1,e})\right] E[S_1], \\ R_2(t) &= E[\lambda_2^+(t - S_{2,e})]E[S_2] = E\left[\sum_{j=1}^{\infty} p^j \lambda(t - S_1^{*j} - S_2^{*j-1} - S_{2,e})\right] E[S_2], \end{aligned} \tag{5.2}$$

Note that the time-lag that exists between the arrival rate function and the offered load function is influenced both by service time and delay time. We shall see later (5.3) that, in some special cases, these offered load expressions have an explicit solution. Nevertheless, in most cases, numerical or approximated solutions are called for. In Section 5.1 we provide a numerical method for calculating $R(t)$, which is applicable when service times are exponentially distributed. When this is not the case we describe two methods for approximating the offered load expression: (a) using Taylor series (5.2) and, (b) in case of periodic arrival rates, using time-series methods (5.3).

5.1 Numerical Approximation of the Offered-Load Measure for General Arrival-Rate Functions with Exponential Service-Time Distribution

In this section, we calculate $R(t)$ as a fluid solution of an infinite-server system, in cases where S_i are exponentially distributed. The Erlang-R model is then a time- and state-dependent Markovian open queueing network. We rely on the mathematical framework of Mandelbaum et al. [53], which provides us with a general solution that is suitable for time-varying arrivals, and time-varying staffing policies. Note that with general time-varying arrival rates, the ODE (Ordinary Differential Equation) system that we develop here is unlikely to be tractable analytically. Nevertheless, we can solve it numerically.

Theorem 2. *If S_i are exponentially distributed then (5.2) is given by the unique solution of the following ODE:*

$$\begin{aligned} \frac{d}{dt}R_1(t) &= \lambda_t + \delta R_2(t) - \mu R_1(t), \\ \frac{d}{dt}R_2(t) &= p\mu R_1(t) - \delta R_2(t). \end{aligned} \tag{5.3}$$

The initial condition is: $R(-\infty) = (0, 0)$.

Proof: see Appendix A.2 on page 149.

This form of the offered load function can be easily solved numerically, using a simple spreadsheet or a mathematical program. We have used this method for calculating $R(t)$ for the experiments in Sections 6 and 7.

5.2 Offered-Load Approximation for General Arrival-Rate Functions with General Service-Time Distribution

In this section, we develop approximate solutions for general arrival rate functions with general service-time distributions. This provides a practical method for calculating the offered load, as well

as insights on the effect of re-entering customers on system's performance. The term $E[\lambda_i^+(t-S_{i,e})]$ in the offered load function (5.1) is difficult to compute because we have a stochastic time-lag within the arrival rate function. Eick et al. [22] solved this problem by approximating the arrival rate using smoothing methods. Approximating $\lambda(t)$ as a polynomial function enables one to express this expectation in terms of moments of $S_{i,e}$. We use the same method here.

5.2.1 Linear Arrival Rate Functions and First-Order Taylor-Series Approximations.

In some environments, $\lambda(t)$ is not constant but has a trend. This can be approximated by a linear function of the following form:

$$\lambda(t) = a + bt, \quad t \geq 0. \tag{5.4}$$

Proposition 1. *For λ linear, as in (5.4),*

$$\begin{aligned} R_1(t) &= \lambda(t - E[S_{1,e}])\frac{E[S_1]}{1-p} - b\frac{p}{(1-p)^2}E[S_1](E[S_1] + E[S_2]). \\ &= \frac{E[S_1]}{1-p}\lambda\left(t - E[S_{1,e}] - \frac{p}{1-p}(E[S_1] + E[S_2])\right). \end{aligned}$$

Proof: see Appendix A.2 on page 150.

Note that $\lambda(t - E[S_{1,e}])\frac{E[S_1]}{1-p}$ is exactly the first-order Erlang-C approximation for $R(t)$, when the arrival rate is multiplied by $\frac{1}{(1-p)}$, which is the average number of visits in a Needy state. When b is positive, the offered load of the Erlang-C model will exceed that of the Erlang-R model, and vice versa. The second representation emphasizes the time-lag between the offered load and the arrival rate. The time-lag is the sum of expected Needy and Content time durations. However, the linear case is not rich enough to separate the amplitude and phase effects. To do this, we later use the sinusoidal case. Nevertheless, these observations already justify the following conclusion:

Conclusion 1. *When the arrival rate is a non-constant linear function, with $b \neq 0$, using Erlang-C model, when customers re-enter the system, will over- or under-estimate the number of servers required, as compared with the corresponding Erlang-R model.*

This conclusion is also based on the connection between the offered load measure and staffing levels, as explained in the introduction.

The linear function results could be generalized to a wider sets of arrival rates using a first-order Taylor-series approximation. This approximation takes the form:

$$\lambda(t-u) = \lambda(t) - \lambda^{(1)}(t)u \quad for \quad u \geq 0, \tag{5.5}$$

where $\lambda^{(1)}(t)$ is the derivative of $\lambda(t)$ evaluated at time t. Using this expression, one develops the first-order approximation of $R_i(t)$ as follows:

Corollary 1. *For linear approximations of $\lambda(t)$, as in (5.5),*

$$\begin{aligned} R_1(t) &= \frac{E[S_1]}{1-p}\lambda(t - E\left[S_{1,e}\right]) - \lambda^{(1)}(t)\frac{p}{(1-p)^2}E[S_1]\left(E\left[S_1\right] + E\left[S_2\right]\right) \\ &= \frac{E[S_1]}{1-p}\lambda\left(t - E\left[S_{1,e}\right] - \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_{2,e}\right]\right)\right). \end{aligned} \tag{5.6}$$

5.2.2 Quadratic Arrival Rate Functions and Second-Order Taylor-Series Approximations.

We now consider a second-order Taylor-series approximation for the arrival-rate function $\lambda(t)$:

$$\lambda(t-u) = \lambda(t) - \lambda^{(1)}(t)u + \lambda^{(2)}(t)\frac{u^2}{2} \quad for \quad u \geq 0, \tag{5.7}$$

where $\lambda^{(k)}(t)$ is the k^{th} derivative of $\lambda(t)$ evaluated at time t. Then, from (5.7) and (5.2) we get the following approximation for $R_1(t)$:

Theorem 3. *For a quadratic approximations of $\lambda(t)$, as in (5.7),*

$$\begin{aligned} R_1(t) = \frac{E[S_1]}{1-p}\Bigg[&\lambda\left(t - E\left[S_{1,e}\right] - \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right) \\ &+\frac{1}{2}\lambda^{(2)}(t)\left(VAR[S_{1,e}] + \frac{p}{1-p}(VAR[S_1] + VAR[S_2]) + \frac{p}{(1-p)^2}\left(E[S_1] + E[S_2]\right)^2\right)\Bigg] \end{aligned} \tag{5.8}$$

Proof: see Appendix A.2 on page 151.

This expression differs from the second-order Taylor-series approximation of Erlang-C, both in time-lag and amplitude. Erlang-C's second-order approximation, as given in Whitt [72], is:

$$R(t) = E[S]\left[\lambda(t - E\left[S_e\right]) + \frac{1}{2}\lambda^{(2)}(t)VAR[S_e]\right].$$

Whitt interpreted the first moment of S_e as the deterministic time-lag between $R(t)$ and $\lambda(t)$, and the second moment as a deterministic magnitude shift.

In Erlang-R, the situation becomes more complex, which will be easier to interpret using the following notation: Let us define M as the number of returns to service; $M \sim Geom_{\geq 0}(1-p)$. Then we can rewrite $R_1(t)$ as follows:

$$\begin{aligned} R_1(t) &= \frac{E[S_1]}{1-p}\Big[\lambda\left(t - E\left[S_{1,e}\right] - E[M]\left(E\left[S_1\right] + E\left[S_2\right]\right)\right) \\ &+\frac{1}{2}\lambda^{(2)}(t)\left(VAR[S_{1,e}] + E[M](VAR[S_1] + VAR[S_2]) + VAR[M]\left(E[S_1] + E[S_2]\right)^2\right)\Big] \\ &= \frac{E[S_1]}{1-p}\Big[\lambda\left(t - E\left[S_{1,e}\right] - E[M]E\left[S_1 + S_2\right]\right) \\ &+\frac{1}{2}\lambda^{(2)}(t)\left(VAR[S_{1,e}] + E[M]VAR[S_1 + S_2] + VAR[M]\left(E[S_1 + S_2]\right)^2\right)\Big]. \end{aligned}$$

The last equality is due to the independence of S_1 and S_2. This representation emphasizes that the deterministic time-lag is not solely a function of $S_{1,e}$, but an average of the total cycles time plus one last residual service. By Wald's identity, the average of total cycles time is $E[M]\left(E\left[S_1\right]+E\left[S_2\right]\right)$, and its variance is $E[M]VAR[S_1+S_2]+VAR[M]\left(E[S_1+S_2]\right)^2$. Therefore, the deterministic magnitude shift is also not only a function of the variance of $S_{1,e}$, but the second moment of the total average cycles length plus the last residual service.

We will demonstrate the effect of these differences in further detail in Sections 6.1 and 5.3.2.

5.3 Analysis of Special Cases and Managerial Insights: The Offered-Load for Sinusoidal Arrival Rate

In this section, we analyze the offered-load expression for the special case of sinusoidal arrival rate. There are two reasons for using the sine function: First, any periodic time-varying arrival rate could be expressed by a combination of sine functions, via time-series methods. Second, using a sine function enables us to compute a closed-form solution to the offered-load function in some special cases which, in turn, reveals mathematically the behavior of the offered load. Therefore, this example will give us a more precise understanding of the role of frequency, service, and delay times in our system.

Define

$$\lambda(t)=\bar{\lambda}+\bar{\lambda}\kappa\sin\left(2\pi t/f\right)=\bar{\lambda}+\bar{\lambda}\kappa\sin\left(\omega t\right),\quad t\geq 0,$$

where $\bar{\lambda}$ is the average arrival rate, κ is the relative amplitude, f is the period, and ω is the frequency. Incorporating this arrival rate into (5.2) yields

$$\begin{aligned}
R_1(t)&=\sum_{j=0}^{\infty}p^jE[S_1]E\left[\bar{\lambda}+\bar{\lambda}\kappa\sin\left(\omega(t-S_{1,e}-S_1^{*j}-S_2^{*j})\right)\right]\\
&=\sum_{j=0}^{\infty}p^jE[S_1]\bar{\lambda}+\sum_{j=0}^{\infty}p^jE[S_1]E\left[\bar{\lambda}\kappa\sin\left(\omega(t-S_{1,e}-S_1^{*j}-S_2^{*j})\right)\right] \qquad (5.9)\\
&=\frac{\bar{\lambda}}{1-p}E[S_1]+E[S_1]\bar{\lambda}\kappa\sum_{j=0}^{\infty}p^jE\left[\sin\left(\omega\left(t-S_{1,e}-S_1^{*j}-S_2^{*j}\right)\right)\right].
\end{aligned}$$

From (5.9) it is obvious that the amplitude of $R_1(t)$ is determined by the infinite sum expression. Using the sine formula $\sin(x-y)=\sin x\cos y-\sin y\cos x$, we get

$$\begin{aligned}
R_1(t)=\frac{\bar{\lambda}}{1-p}E[S_1]+E[S_1]\bar{\lambda}\kappa\sum_{j=0}^{\infty}p^jE\Big[&\sin\left(\omega(t)\right)\cos\left(\omega(S_{1,e}+S_1^{*j}+S_2^{*j})\right)\\
&-\sin\left(\omega(S_{1,e}+S_1^{*j}+S_2^{*j})\right)\cos\left(\omega(t)\right)\Big]. \qquad (5.10)
\end{aligned}$$

The same analysis could be performed by examining $\lambda^+(t)$ over time.

$$\begin{aligned}\lambda_1^+(t) &= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j E\left[\sin\left(\omega\left(t - S_1^{*j} - S_2^{*j}\right)\right)\right] \\ &= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j E\left[\sin\left(\omega(t)\right)\cos\left(\omega(S_1^{*j} + S_2^{*j})\right) - \sin\left(\omega(S_1^{*j} + S_2^{*j})\right)\cos\left(\omega(t)\right)\right].\end{aligned} \tag{5.11}$$

Substituting $\lambda^+(t)$ into (5.1) will form the same $R(t)$ expressions. Using $\lambda^+(t)$, we see that the amplitude of the total arrival rate is determined by the expression $\sum_j p^j E[E[\sin(\omega(t - S_1^{*j} - S_2^{*j}))]]$.

5.3.1 Exponential Service Times

We will now analyze (5.10) for the case of exponential service times. We assume that $S_1 \sim exp(\mu)$ and $S_2 \sim exp(\delta)$.

Theorem 4. *Assuming S_i is exponentially distributed, (5.10) has the following form:*

$$R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sqrt{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}\cos\left(\omega t + \pi + tan^{-1}(\theta)\right)$$

where

$$\theta = i \cdot \frac{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} + \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} - \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}} = \frac{-\mu(-\delta^2 + p\delta^2 - \omega^2)}{\omega(\delta^2 + \omega^2 + p\mu\delta)}.$$

Proof: see Appendix A.3 on page 152.

Therefore, the amplitude of $R_1(t)$ is given by

$$Amp(R_1(t)) = \bar{\lambda}\kappa\sqrt{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}$$

and the phase shift of $R_1(t)$ with respect to the entrance arrival rate is given by

$$Phase(R_1(t)) = \frac{1}{2\pi} cot^{-1}\left(\frac{\mu(\delta^2 - p\delta^2 + \omega^2)}{\omega(\delta^2 + \omega^2 + p\mu\delta)}\right)$$

The same expansion could be performed for $\lambda_1^+(t)$:

Theorem 5. *Assuming that S_i is exponentially distributed, (5.11) has the following form:*

$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sqrt{\frac{(\mu - i\omega)(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{(\mu + i\omega)(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}\cos\left(\omega t + \pi + tan^{-1}(\theta)\right)$$

where

$$\theta = i \cdot \frac{\frac{(\mu - i\omega)(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} + \frac{(\mu + i\omega)(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}{\frac{(\mu - i\omega)(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} - \frac{(\mu + i\omega)(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}} = \frac{\omega^2\delta^2 + \omega^4 + \omega^2 p\mu\delta + \mu^2\delta^2 - \mu^2 p\delta^2 + \mu^2\omega^2}{\mu\omega p\delta(\mu + \delta)}.$$

Proof: see Appendix A.3 on page 154.

Therefore, the amplitude of $\lambda_1^+(t)$ is given by

$$Amp(\lambda_1^+(t)) = \bar{\lambda}\kappa\sqrt{\frac{(\mu - i\omega)(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{(\mu + i\omega)(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}},$$

and the phase shift of $\lambda_1^+(t)$ with respect to the entrance arrival rate is given by

$$Phase(\lambda_1^+(t)) = \frac{1}{2\pi} cot^{-1}\left(\frac{\omega^2\delta^2 + \omega^4 + \omega^2 p\mu\delta + \mu^2\delta^2 - \mu^2 p\delta^2 + \mu^2\omega^2}{\mu\omega p\delta(\mu + \delta)}\right)$$

Note that there is a simple relation between the amplitudes of $R(t)$ and $\lambda_1^+(t)$.

$$Amp(R_1(t)) = Amp(\lambda_1^+(t))\sqrt{\mu^2 + \omega^2}.$$

This relation separates two influences on the offered-load amplitude: $Amp(\lambda_1^+(t))$ represents the influence of returning customers, and $\sqrt{\mu^2 + \omega^2}$ the influence of the last service.

To further investigate the relative amplitude of the offered load ($R_1(t)$) and the aggregate arrival rate ($\lambda_1^+(t)$), we state the following proposition that highlights some of the limits of $R_1(t)$ and $\lambda_1^+(t)$ with respect to ω and δ:

Proposition 2. *In the case of sinusoidal arrival rates and exponential service times, if μ and δ are fixed, it follows that:*

$$\lim_{\omega\to 0} R_1(t) = \lim_{\omega\to 0} \frac{E[S_1]\bar{\lambda}}{1-p} + E[S_1]\frac{\bar{\lambda}}{\mu(1-p)}\kappa\sin(\omega t),$$
$$\lim_{\omega\to \infty} R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p},$$
$$\lim_{\omega\to 0} \lambda_1^+(t) = \lim_{\omega\to 0} \frac{\bar{\lambda}}{1-p} + \frac{\bar{\lambda}}{1-p}\kappa\sin(\omega t),$$
$$\lim_{\omega\to \infty} \lambda_1^+(t) = \lim_{\omega\to \infty} \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sin(\omega t),$$

and if μ and ω are fixed,

$$\lim_{\delta\to 0} R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\mu^2+\omega^2}\left(\mu\sin(\omega t) - \omega\cos(\omega t)\right),$$
$$\lim_{\delta\to \infty} R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{(1-p)^2\mu^2+\omega^2}\left((1-p)\mu\sin(\omega t) - \omega\cos(\omega t)\right).$$

Proof: see Appendix A.3 on page 155.

We will use a numerical example to demonstrate the relative amplitude behavior. Figure 4 shows the amplitude of $R_1(t)$ and $\lambda_1^+(t)$ with respect to the amplitude of $\lambda(t)$ (which is $\bar{\lambda}\kappa$), as a function of ω (i.e., when μ and δ are fixed). We observe that, in the range $(0, \infty)$ the relative amplitude of

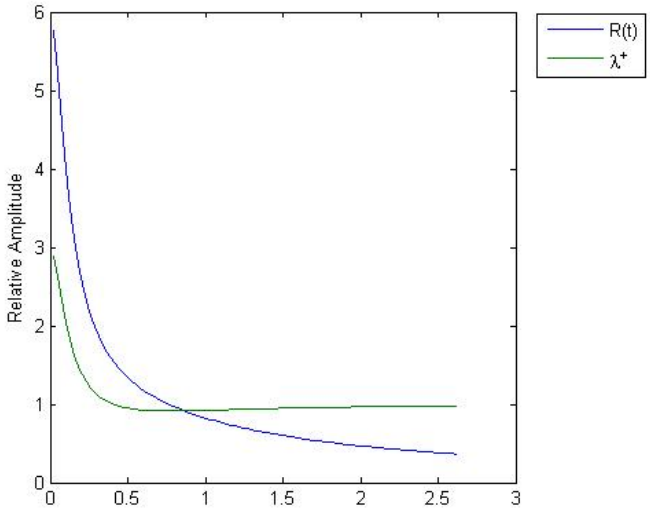


Figure 4: Plot of relative amplitude of $R_1(t)$ and $\lambda_1^+(t)$ with respect to ω

$R_1(t)$ is a decreasing function of ω, starting from the value $\frac{1}{\mu(1-p)}$, and decreasing to 0 as $\omega \to \infty$. On the other hand, $\lambda_1^+(t)$ starts from the value $\frac{1}{1-p}$, and tends to 1 as $\omega \to \infty$. Figure 5 shows the amplitude of $R_1(t)$ with respect to the amplitude of $\lambda(t)$, as a function of ω and δ (when $\mu = 0.5$). We observe that, in the range $(0, \infty)$, the relative amplitude of $R_1(t)$ is an increasing function of δ, starting from the value $\frac{1}{\sqrt{\mu^2+\omega^2}}$, and increasing to $\frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\sqrt{(1-p)^2\mu^2+\omega^2}}$ as $\delta \to \infty$. When $\delta \to 0$ the extreme values of $R_1(t)$ are $\max_t(R_1(t)) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\sqrt{\mu^2+\omega^2}}$. When $\delta \to \infty$ the extreme values of $R_1(t)$ are $\max_t(R_1(t)) = \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{\bar{\lambda}\kappa}{\sqrt{(1-p)^2\mu^2+\omega^2}}$.

Figure 6 shows the phase shift of $R_1(t)$ from $\lambda(t)$ as a function of ω. This phase shift is the sum of two phases: the phase between $R_1(t)$ and $\lambda_1^+(t)$, and the phase between $\lambda_1^+(t)$ and $\lambda(t)$. One is due to returning customers, and the other is due to the last service.

5.3.2 Comparison to Erlang-C

In this section, we compare the amplitude and phase shift of the offered-loads for Erlang-C with those of the Erlang-R model.

The amplitude of the offered-load in Erlang-C is given by: $Amp(R^c(t)) = \frac{\bar{\lambda}\kappa}{\sqrt{\mu^2+\omega^2}}$, and its phase shift is $\theta^c = \frac{1}{2\pi} cot^{-1}(\mu/\omega)$ (See [21]). We compare the Erlang-R model to an Erlang-C model with concatenated services (i.e. service rate equals $(1-p)\mu$), and an arrival rate of $\lambda(t)$. We call this

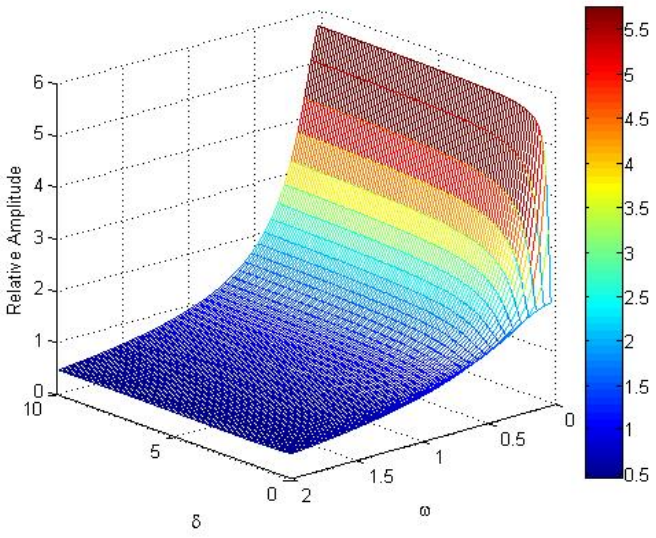


Figure 5: Plot of relative amplitude of $R_1(t)$ with respect to δ and ω

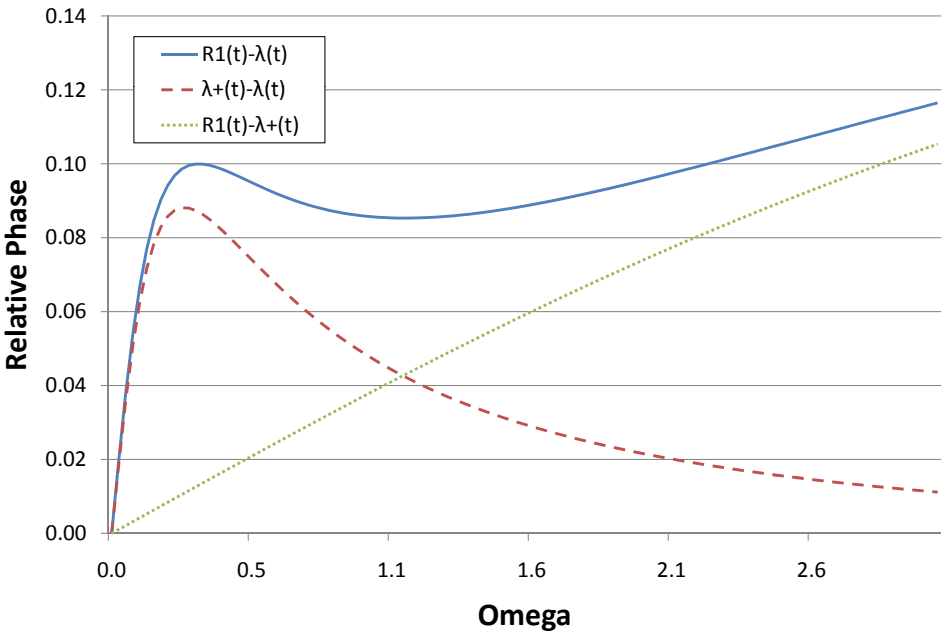


Figure 6: The relative phase between $R_1(t)$ and $\lambda(t)$ as a function of ω

Erlang-C model a *multi-service* Erlang-C. The ratio between the amplitudes is given by

$$AmpRatio = \frac{Amp(R_1(t))}{Amp(R^c(t))} = \frac{\bar{\lambda}\kappa\sqrt{\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}}{\frac{\bar{\lambda}\kappa}{\sqrt{((1-p)\mu)^2+\omega^2}}}$$
$$= \sqrt{\frac{\delta^2+\omega^2}{((\mu-i\omega)(\delta-i\omega)-p\mu\delta)((\mu+i\omega)(\delta+i\omega)-p\mu\delta)}}\Big/\frac{1}{\sqrt{((1-p)\mu)^2+\omega^2}}$$

and the ratio between the phase shifts is given by

$$\frac{Phase(R_1(t))}{Phase(R^c(t))} = \frac{cot^{-1}\left(\frac{-\mu(-\delta^2+p\delta^2-\omega^2)}{\omega(\delta^2+\omega^2+p\mu\delta)}\right)}{cot^{-1}\left(\frac{(1-p)\mu}{\omega}\right)} = \frac{\left(\pi+2tan^{-1}\left(\frac{\mu\left(-\delta^2+p\delta^2-\omega^2\right)}{\omega(\delta^2+\omega^2+p\mu\delta)}\right)\right)}{\left(\pi-2tan^{-1}\left(\frac{(1-p)\mu}{\omega}\right)\right)}.$$

Theorem 6. *When the arrival rate is sinusoidal and the service times are exponentially distributed:*

1. *The amplitude of the offered-load under Erlang-R model is smaller than the amplitude of the offered-load under* multi-service *Erlang-C model, for every set of parameters.*

2. *The amplitude ratio gets its minimal value when* $\omega = \sqrt{\delta\mu(1-p)}$.

Proof: see Appendix A.4 on page 157.

The first part of the theorem states that the amplitude ratio is smaller than one, and implies that returning customers have a stabilizing effect on the system. An example of the difference between the amplitudes is given in the left diagram of Figure 7. Having a smaller amplitude means that for some part of the cycle, $R_1(t)$ is higher, and in the other part $R^c(t)$ will be higher, as shown in Figure 8. The implication will be that Erlang-C will over- or under- staff. To understand the impact of this analysis on service level, refer to Section 6. The second part of the theorem identifies the cases in which the difference between the amplitudes is higher. Note that the EW environment is one in which the parameters of p, μ, δ and ω are such that the ratio is close to its minimal value.

The phase ratio as a function of ω (see the right diagram of Figure 7) is larger than one up to $\omega = \sqrt{\frac{2\delta^2+p\delta\mu-p^2\delta\mu}{p}}$, and from that point onward it is smaller than one. Therefore, for certain values of ω, the Erlang-C offered load leads the Erlang-R offered load and for other values it lags behind, as shown in Figure 9.

From Theorems 2 and 6 we can gain an understanding when the influence of the returning customers is most significant and thus requires the use of the Erlang-R model.

Corollary 2. *When the arrival rate is sinusoidal and the service times are exponentially distributed, if* $\omega \searrow 0$, $\omega \nearrow \infty$, *or* $\delta \nearrow \infty$ *the difference between the offered-load of Erlang-R and Erlang-C becomes negligible.*

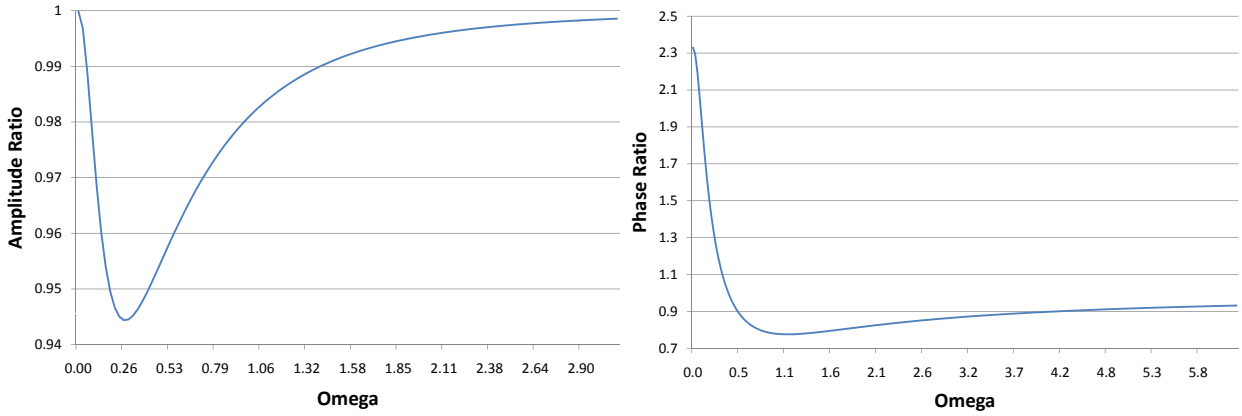


Figure 7: The ratio of amplitudes and phases between Erlang-R and Erlang-C as a function of ω

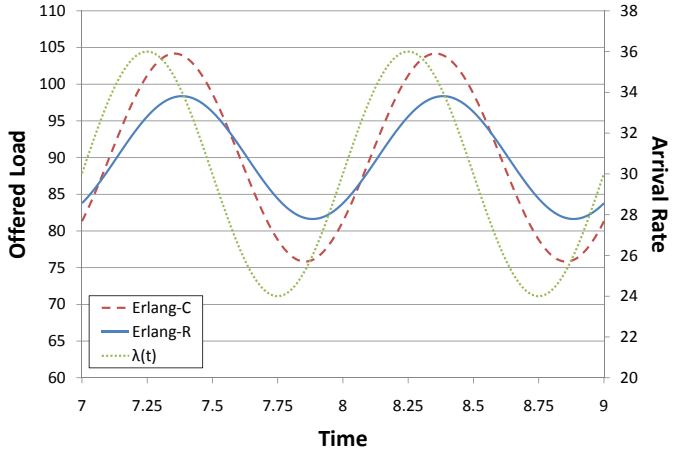


Figure 8: Example: Erlang-C under- or over-estimates the offered load

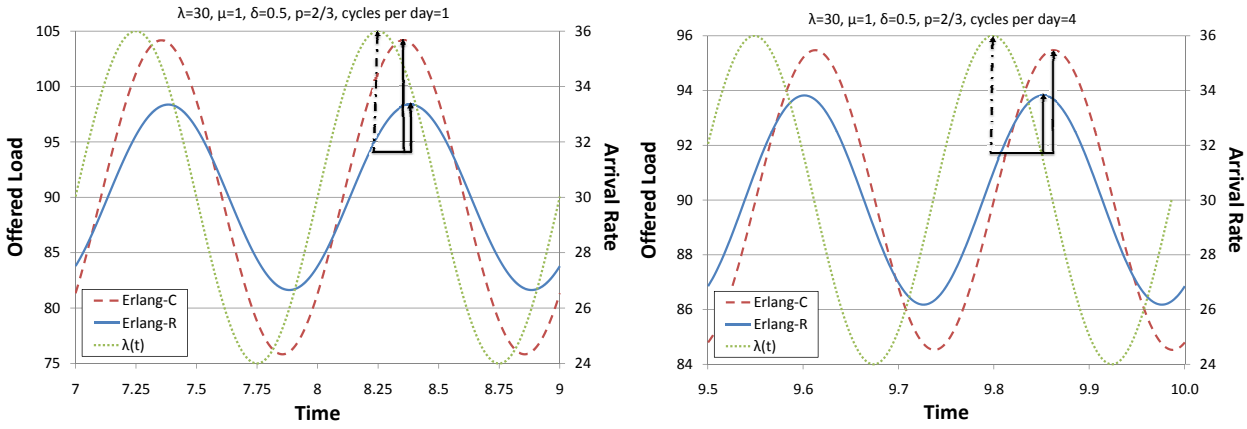


Figure 9: Example: Erlang-C under- or over-estimates the time-lag

An intuitive explanation for this finding is that when ω tends to infinity, the arrival rate changes so slowly that the system reaches a steady-state. In this case, the offered-load becomes constant; this is true for both the Erlang-C and Erlang-R models. When ω tends to zero, the arrival rate changes so rapidly that its changes are assimilated in the variance of the arrival process. As δ tends to infinity, customers immediately return to the Needy state; thus the system behaves as if the services were concatenated.

6 Validation of MOL Staffing

We now propose a staffing procedure for the Erlang-R model. In Section 4 we showed that the steady-state marginal distribution of the Needy state is identical both in Erlang-C and Erlang-R models. Therefore, if the arrival rate is constant, and the system is in a steady state, the staffing recommendations will also be equal. There is thus no need to use the more complex Erlang-R model to determine staffing in this case. More precisely, if we want the system to operate in the QED regime, we should use the square-root formula $s = R + \beta\sqrt{R}$, where R is the offered load given by $R = \frac{\lambda}{1-p}E[S_1]$ and β is chosen according to the Halfin-Whitt formula [39]. On the other hand, if the arrival rate is time varying, there is a difference between the two models. For time varying environments, we propose the use of the MOL approximation (e.g. Massey and Whitt [60]). We will compare it to two other approaches: time-varying Erlang-C model and the PSA (Piecewise Stationary Analysis) approximation.

The MOL Algorithm for Erlang-R runs as follows:

1. Calculate the time-varying offered load $R(t)$; in the case of exponential service times simply solve the differential equations (5.3).

2. Staff according to the square-root formula: $s(t) = R(t) + \beta\sqrt{R(t)}$, where β is chosen according to the steady-state Halfin-Whitt formula.

The second stage takes place because both the Erlang-R and Erlang-C have the same steady-state marginal distribution for the Needy state.

We use simulation to validate this approach. In the first case study we consider sinusoidal arrival rates in a large system and, in the second, a small system with an arrival-rate shape that is typical of hospitals. We now describe each of the case studies.

6.1 Case Study 1 - Large System; Sinusoidal Arrival Rates; Exponential Service Times

In this case study, we validate our assumption (based on Feldman et al. [26]) that the MOL algorithm stabilizes system's performance over time. The main performance measure we consider is the probability of waiting ($P(W > 0)$), but other performance measures are also considered. We will use a stylized arrival rate with a sinusoidal shape. To this end, define the following arrival rate function:

$$\lambda_t = \bar{\lambda} + \bar{\lambda}\kappa sin(2\pi t/\psi) = \bar{\lambda} + \bar{\lambda}\kappa sin(\omega t),$$

where $\bar{\lambda}$ is the average arrival rate, κ is the relative amplitude, and ψ is the period length (ω is the frequency).

We use relatively large $\bar{\lambda}$ since we start our validation process in a large system, where the asymptotic approximations are expected to work well.

The parameters of this experiment are: $\bar{\lambda} = 30$ customers per hour, $p = 2/3$, $\kappa = 0.2$, $\psi = 24$ hours, $\mu = 1$, $\delta = 0.5$, and $\beta \in 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9, 1.0, 1.5$. Figure 10 shows the behavior of the fluid solution (5.3) over time. It presents the arrival rate ($\lambda(t)$), the aggregated arrival rate ($\lambda_1^+(t)$), the numerical solution of the offered load in Needy state ($R_1(t)$), and the recommended staffing when $\beta = 0.2$. Note that $p = 2/3$ means that the average number of cycles for each customer is three. Therefore, we can see clearly that $\lambda_1^+(t)$ varies around 90, as expected by the expression (5.11). We also observe the time-lag that exists between the arrival rate and the offered load.

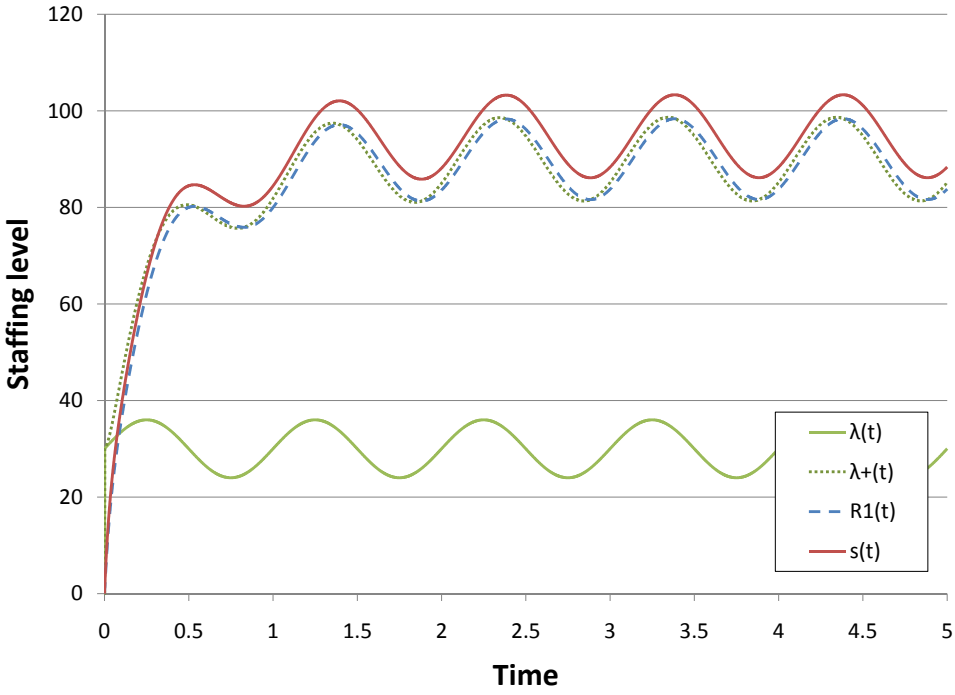
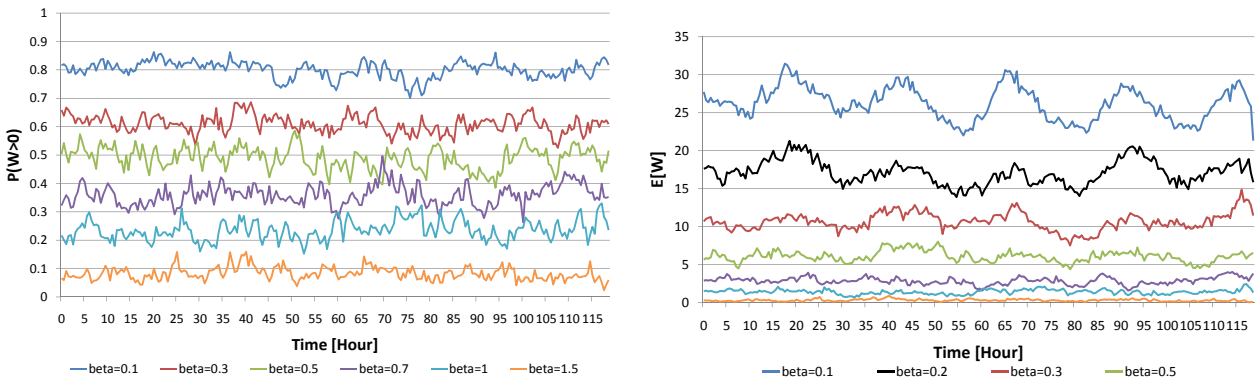


Figure 10: Case study 1 - Arrival rate, offered-load, and staffing

We have calculated $s(t)$ according to the square-root formula: $s(t) = R(t) + \beta\sqrt{R(t)}$, and rounded the results. We used one hundred replications for each β value. The left diagram in Figure 11 shows the performance measure $P(W > 0)$ over time for various values of β. We see that the performance measure is stable, which indicates that the MOL algorithm works well. The left diagram in Figure 12 shows the changes in servers' utilization over time for each value of β. This performance measure is also stable. Thus our staffing procedure not only stabilizes the service level

but also server utilization. In the right diagram of Figure 12, we compare the average utilization over time with the theoretical values. The latter were calculated using the steady-state solution of our model, when given average values of λ and s. We see that the two are almost identical. The right diagram in Figure 11 shows the performance measure E[W] over time for various values of β. We note that, as β grows, this measure becomes more stable. We also see that each value of β results in a different average value of E[W], and that the relative order between these values matches the order of β values.

Figure 13 shows the conditional distribution of the waiting time given delay $(W|W > 0)$, for three values of β (0.1,0.5, and 1.4). We compare them to the steady-state theoretical distribution, which is exponential with rate $n\mu(1-\rho)$ (as stated in Theorem 1). The simulation results depict the distribution of waiting times from all replication, over the entire time horizon. We observe a very good fit for $\beta = 0.5$ (QED) and $\beta = 1.4$ (QD (Quality Driven)), but when β is small (0.1-ED (Efficiency Driven)), the theoretical distribution does not match well the simulation results. This is in line with our observations for E[W], where small values of beta give rise to performance that varies in time and thus does not correspond to steady-state. Figure 14 shows the performance measure $P(W > T)$ (the probability to wait more than T units of time) over time, for various values of β. We used a value of $T = 5$ minutes. We note that again the performance measure is stable.

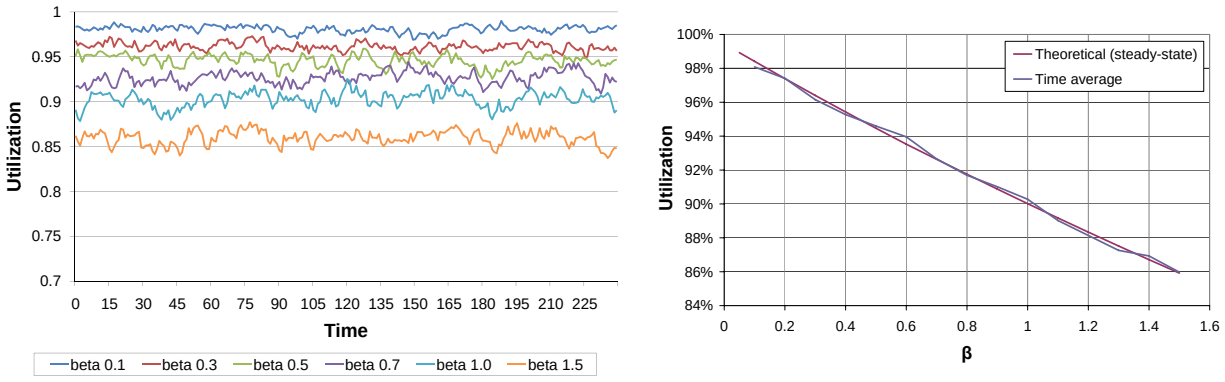


$P(W > 0)$ $E[W]$

Figure 11: Case study 1 - Simulation results for various β values in large systems

As mentioned before, we expect the relation between $P(W > 0)$ and β to fit the Halfin-Whitt formula. We have validated this by calculating the average waiting probability over time for each value of β, and compared it to the Halfin-Whitt formula [39]. In Figure 15, the two are very similar to each other.

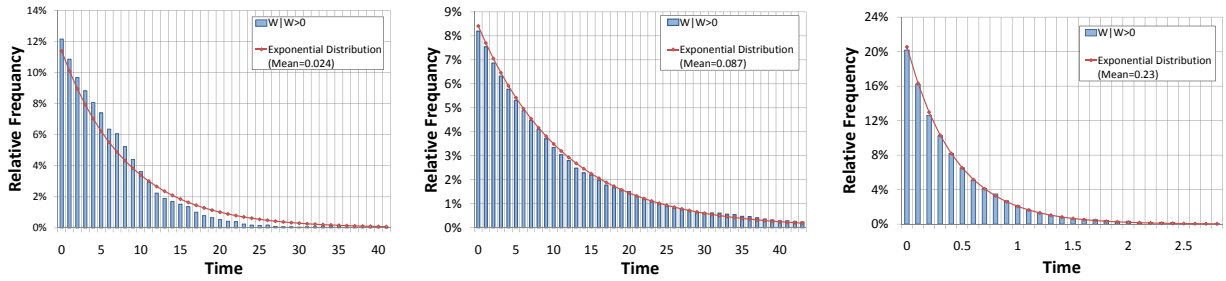
We conclude this case study with the following



utilization for various β values

time averages vs. theoretical values

Figure 12: Case study 1 - Simulation results of server utilization



$(\beta = 0.1)$ $(\beta = 0.5)$ $(\beta = 1.4)$

Figure 13: Case study 1 - A comparison of the histogram of $W|W > 0$ with the corresponding theoretical distribution

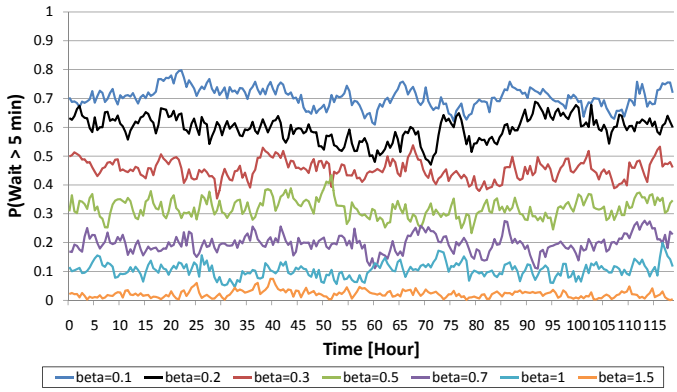


Figure 14: Case study 1 - Simulation results of $P(W > T)$ for various β values in large systems

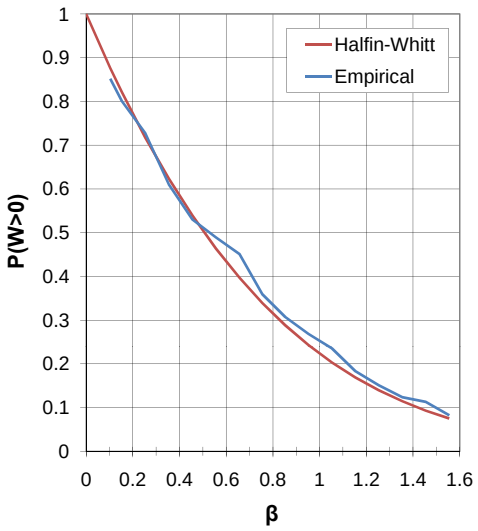


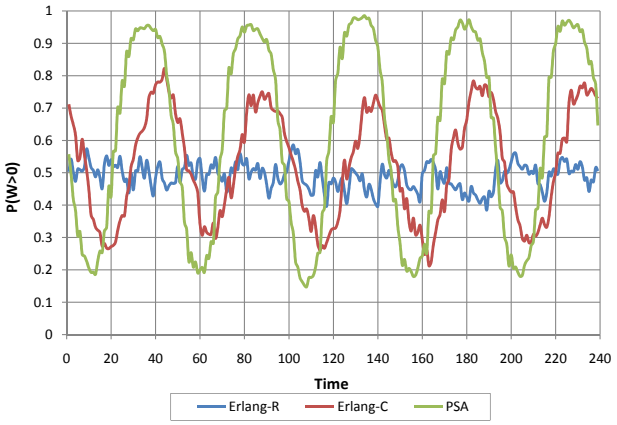
Figure 15: Case study 1 - A comparison of the Halfin-Whitt formula to the simulation results

Conclusion 2. *For a large enough system in the QED regime* ($\beta > 0.3$)*, the MOL approach stabilizes all performance measures. Consequently, any pre-specified QED service level can be achieved stably over time.*

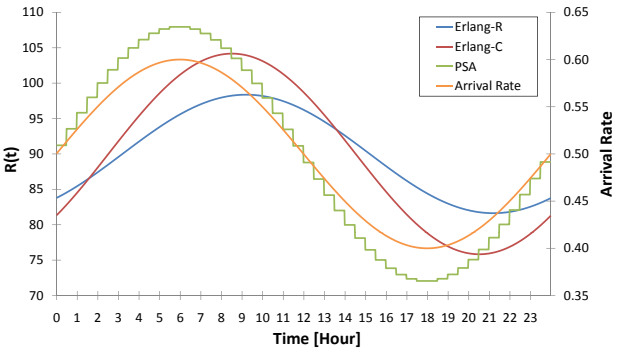
In many applications, researchers use the Erlang-C to model systems in which customers return multiple times for service. For example, Green [33, 34] used the Lag-SIPP approach based on M/M/s (Erlang-C) for staffing doctors at an EW. We would like to compare the outcome of using Erlang-R staffing against that of using Erlang-C staffing, the latter based on one of two methods: MOL and PSA. (We are using MOL since it is known to work very well for Erlang-C [26].) The performance measure we focus on is the delay probability, setting its target level to 0.5 (hence $\beta = 0.5$). The left diagram of Figure 16 shows that using Erlang-R stabilizes the probability of waiting, but the use of Erlang-C does not. As one can anticipate, using a simpler method such as PSA is even worse, resulting in a less stable system. This is because PSA uses the following approximation for the offered load: $R(t) = \bar{\lambda}(t)\frac{E[S]}{1-p}$. PSA staffing does not take into account either the time-lag or the reentrant effects.

The differences in performance have a very simple explanation, when one considers the offered-load function $R(t)$, calculated for each method (see the right diagram of Figure 16). We observe that for one half of the cycle, Erlang-C will over- estimate $R(t)$, resulting in over staffing which, in turn, results in a better performance than specified. However, in the other half cycle, the opposite occurs, causing the performance to be worse than specified. Erlang-R, in contrast, stabilize performance

over the whole horizon.



$P(W > 0)$



$R(t)$

Figure 16: Case study 1 - Comparison between Erlang-R, Erlang-C, and PSA

6.2 Case Study 2 - Small System; Hospitals' Arrival Rates

In the second case study, we investigate the use of the MOL algorithm in small systems, specifically in setting staffing levels for EW physicians. We also take a more complex arrival rate, namely the actual arrival rate function of the Emergency Ward from Figure 17. Values for p, μ, and δ were also inferred from that EW data, as will be articulated in the sequel. Our goal, as before, is to verify whether staffing according to the MOL algorithm stabilizes performance over time.

There are obvious problems in applying our MOL approach in small systems. First, our approximations are expected to be less accurate, being limits as systems grow indefinitely. (In our simulation, the number of servers changes between one and eight.) Second, rounding a "theoretical" need of 1.5 servers up to two servers means adding 30% excess capacity to the required capacity, which suggests difficulties in stablizing performance around pre-specified values. Relate to this is the fact that the set of feasibly performance measures is manifestly discrte for small systmes: changing staffing level of a small system by a single server could significantly change its performance. Finally, one can not have an EW operate with no doctors, and for small servers this lower bound of 1 plays a binding role. It is therefore unclear whether, under these circumstances, we shall still be able to stabilize the systems' performance around a predetermined value. We do show, nevertheless, that it is possible to stablize even such small systems, given specific (though not all, which is expected) target performance levels.

The parameters of this experiment are: $\bar{\lambda} = 9$ customers per hour, $p = 0.69697$, $\mu = 10.9$,

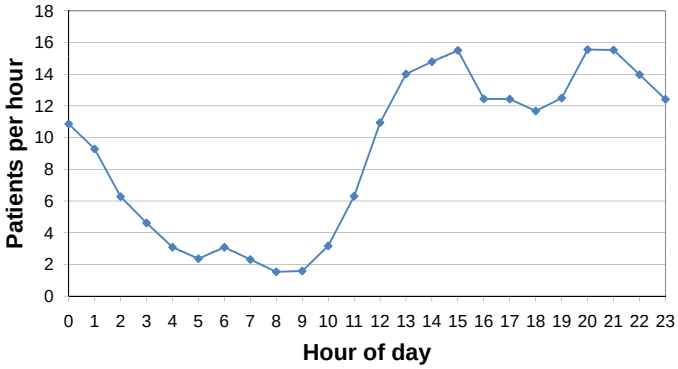


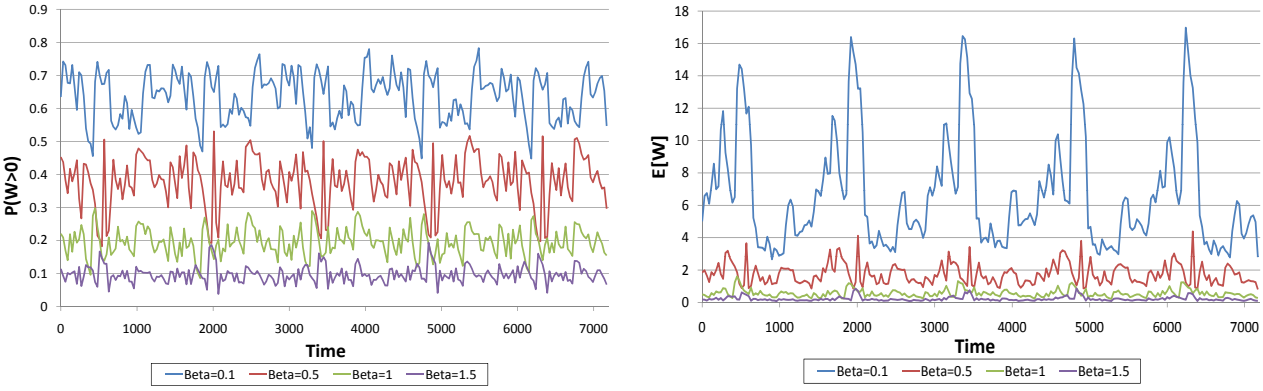
Figure 17: Case study 2 - Plot of arrival rate in emergency ward

β range	s	$P(W > 0)$
$[0, 0.474]$	3	82.4%
$(0.474, 1.055]$	4	34.0%
$(1.055, 1.658]$	5	11.4%
$(1.658, 2.261]$	6	3.0%
1.658 and up	7	0%

Table 1: An example of discrete $P(W > 0)$ as a function of β in small systems

$\delta = 2.3$, and $\beta \in \{0.1, 0.5, 1.0, 1.5\}$. The reason to select only four values of β is that, as stated before, in such a small system one cannot achieve all values of $P(W > 0)$. For example, if $\lambda = 9$ and is constant over time, the offered load is $R = \frac{\lambda}{(1-p)\mu} = \frac{9}{(1-0.697)10.9} = 2.75$. The values $P(W > 0)$ can then have are shown in Table 1. No pure staffing strategy can obtain values between the values shown in the table. A randomized staffing policy can reach other values, but it is not a valid option from a practical point of view, and does not agree with our primary goal that all customers enjoy the same service level at all times.

As before, we calculated $s(t)$ according to the square-root formula, rounded the results to the nearest integer, and used one hundred replication for each β value. The left diagram in Figure 18 shows the performance measure $P(W > 0)$ over time, for various values of β. We see that, the performance measure is relatively stable, and that the four scenarios are separable. Therefore, we conclude that the MOL algorithm works well even in very small systems. The right diagram in Figure 18 shows the performance measure E[W] over time for various values of β. We see that, for very small βs, such as 0.1, E[W] is not stable; but for larger values it is. In addition, we observe that the four scenarios are again separable.



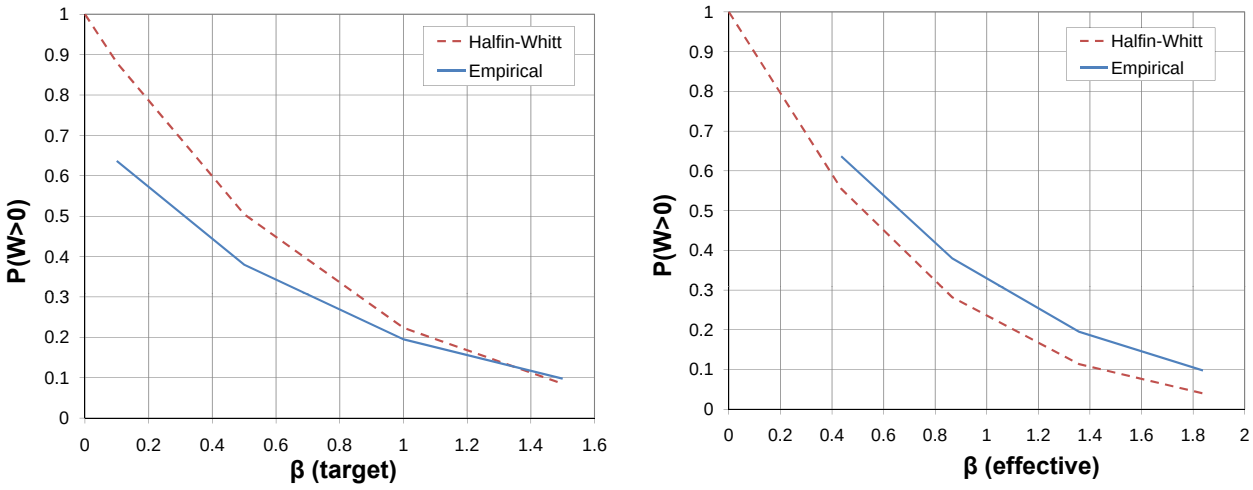
$P(W > 0)$ $E[W]$

Figure 18: Case study 2 - Simulation results for various β values in small systems

As before, we seek to verify whether the relation between $P(W > 0)$ and β fits the Halfin-Whitt formula. The left diagram in Figure 19 shows the relation between these functions when we consider the target β values used in the square-root formula. We see that, in most cases, the empirical function is shifted downwards, and that the gap between the two is reduced as β grows. This is mainly due to the rounding procedure. When β is very small (0.1), the difference between the target β and the one actually used is very large. For example, the average β used, in the 0.1 case, is 0.435; it is four times greater than the value we planned. In this case, therefore, we get much better service performance than the planned performance. In order to eliminate this effect, we look at the relation between these functions when we consider the effective β values actually used, after the rounding process. This is shown in the right diagram in Figure 19. We see that the two functions have the same shape but that the empirical function is shifted upwards. The gap between them appears to be constant. This seems to be the effect of using asymptotic approximations in such a small system. The practical guideline that can be derived from these graphs is that when one targets a specific $P(W > 0)$ value, s/he should use a smaller value of β. One can use the left diagram to derive the exact value. More research is needed in order to find the Halfin-Whitt function for small systems while also considering the rounding effect. As a first step, one can develop such graphs using a steady-state simulation of an Erlang-C model. We can conclude as follows:

Conclusion 3. *The MOL approach stabilizes most performance measures of small systems. In order to achieve a pre-specified service level, one should use a smaller value of β, smaller than that specified by the Halfin-Whitt formula.*

As before, we again compare the performance of Erlang-C against Erlang-R. Figure 20 shows that



(target β values) (effective β values)

Figure 19: Case study 2 - Comparison of the Halfin-Whitt formula to simulation results

there is a clear difference between the performance of Erlang-R and Erlang-C staffing procedures. Erlang-R performances are significantly more stable, even in small systems.

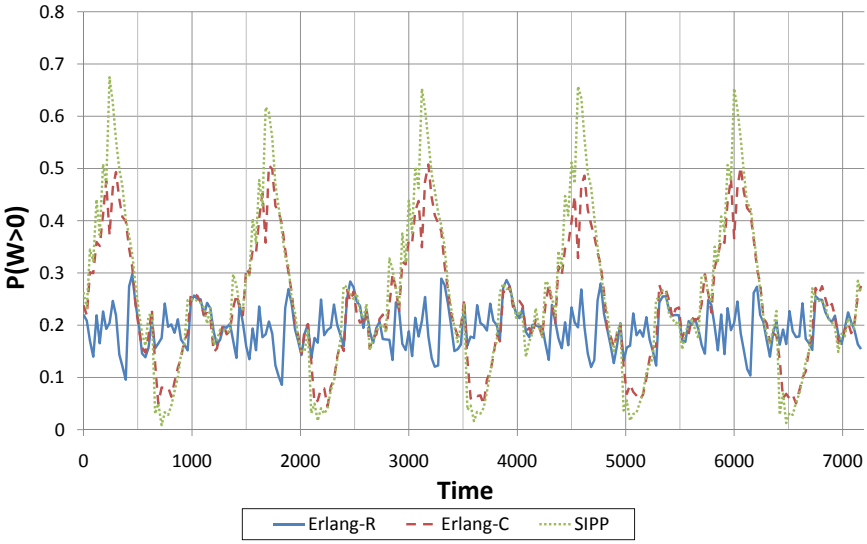


Figure 20: Case study 2 - Plot of $P(W > 0)$ when using Erlang-R and Erlang-C in small system

7 Using Erlang-R for Staffing EW Physicians: Fitting a Simple Model to a Complex Reality

In the following case study, we check the possibility of using an Erlang-R model in planning a real system. We will show that one can use the Erlang-R model for doctors' staffing in an EW, although the real system is much more complicated than our model. Specifically, it is obvious that some of our main assumptions do not hold in this environment. For example, service times are need not be exponentially distributed, and could depend on the load in the EW, as follows from [5, 20]; in particular, service rates need not be constant over the day. In this experiment we used a very accurate and detailed EW simulation model which was developed by Marmor and Sinreich [58]. The simulation is flexible in that it is easily adapted to a given EW. We fit the simulator to the EW of a partner Israeli hospital, and then used the simulator as an accurate portrait of the complex EW reality.

In our EW simulation, there are seven types of patients that enter the EW, and each one goes through a different routing process during their stay. The doctors are divided into four groups, according to their expertise. There is an explicit connection between patients' type and doctors' group. We simplify this complex system into Erlang-R by setting, for each doctor and patient type separately, the parameter values as follows:

- Arrival rate (λ) is the average arrival rate for each hour of the day, as shown in Figure 17.
- Needy times ($\mu = \frac{1}{E[S_1]}$): $E[S_1]$ is estimated by averaging all services given by a specific doctor group to a specific patient's type.
- Content times ($\delta = \frac{1}{E[S_2]}$): $E[S_2]$ is the average time between successive visits of each patient to the doctor.
- Probability of returning to the doctor for an additional service (p): This parameter is calculated from the average number of visits of each patient to their doctor which we take to be $\frac{1}{1-p}$.

We calculated the offered load using the differential equations (5.3), and checked the staffing recommendation in that simulation. Note that in this simulation, the doctors' work is divided between direct and indirect work. The indirect work is work not performed in front of the patient, and has lower priority. We calculated the offered load of the total work. We assumed that changes in staffing could be implemented in a one-hour resolution. For each interval we calculated the average

number of doctors needed and rounded up to the nearest integer. We used one replication of one hundred weeks; the first week was used as setup time, and was not considered in the results.

Figure 21 shows the number of doctors during the day, for each type of doctor, when β equals 0.5. We see that in this small system, the number of doctors varies between one and four. In hospitals, of course, the minimal number of doctors is one (and not zero) since the EW should not have a time without a doctor present. We also observe that the staffing function lags behind the arrival rate function, with an approximate time-lag of two hours.

This system is a very small one. The result is the phenomenon shown in the left diagram of Figure 22, which depicts the probability of waiting for four values of beta: 0.1, 0.5, 1.0, and 1.5. We see that the four cases are not always clearly separable. This is typical of small systems, as we saw in Section 6.2. For example, one can see in Figure 23 that the dips in $P(W > 0)$ when $\beta = 0.1$ are due to times in which the constraint of having at least one doctor of each type present is enforced. In these times, the service levels are not those we aimed for, but the worst possible case. The right diagram of Figure 22 shows the changes in E[W] over time for each β. We see that, as we saw in Case Study 2, $E[W]$ is not stable for $\beta = 0.1$, but is stable for $\beta \geq 0.5$.

We conclude that despite the simplicity of the Erlang-R model, it captures the important aspects of patients' visits in the EW. Using an Erlang-R model, hospital managers can calculate the recommended staffing for doctors. This is not always simple to implement since the next stage requires taking shifts into consideration [61]. Note that the same structure could also be used when considering nurses' work in the EW. It is worthwhile noting that staffing physicians is more challenging than staffing nurses: the latter gives rise to a larger number of servers, hence MOL is expected to be more accurate.

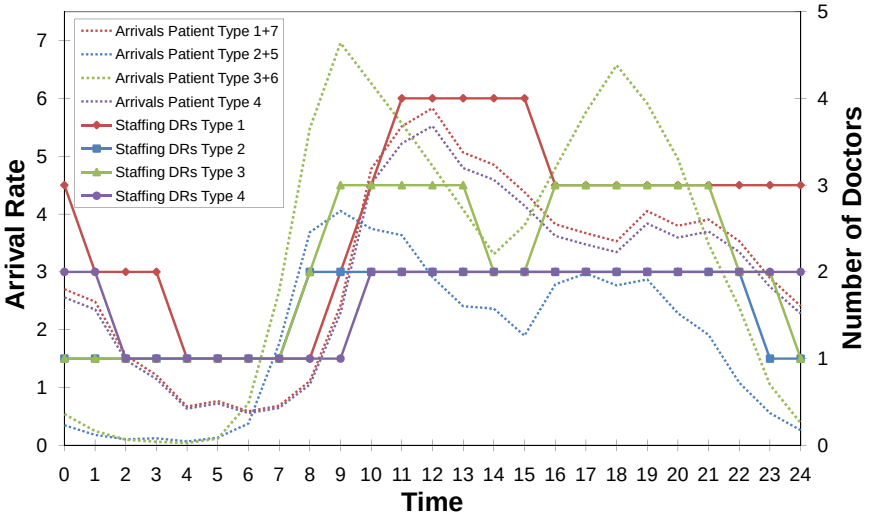
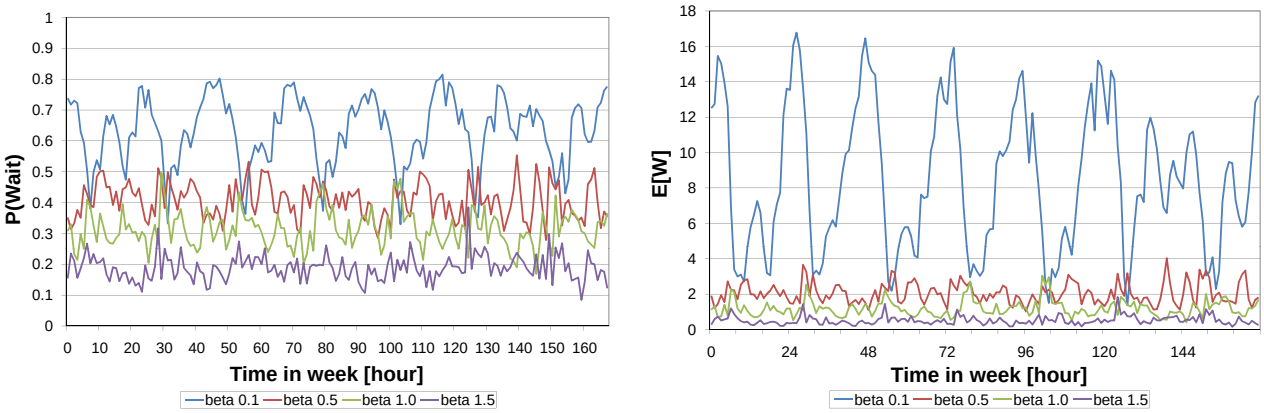


Figure 21: EW case study - Arrivals and Staffing for various β values in EW simulation



$P(W > 0)$ $E[W]$

Figure 22: EW case study - Simulation results for various β values

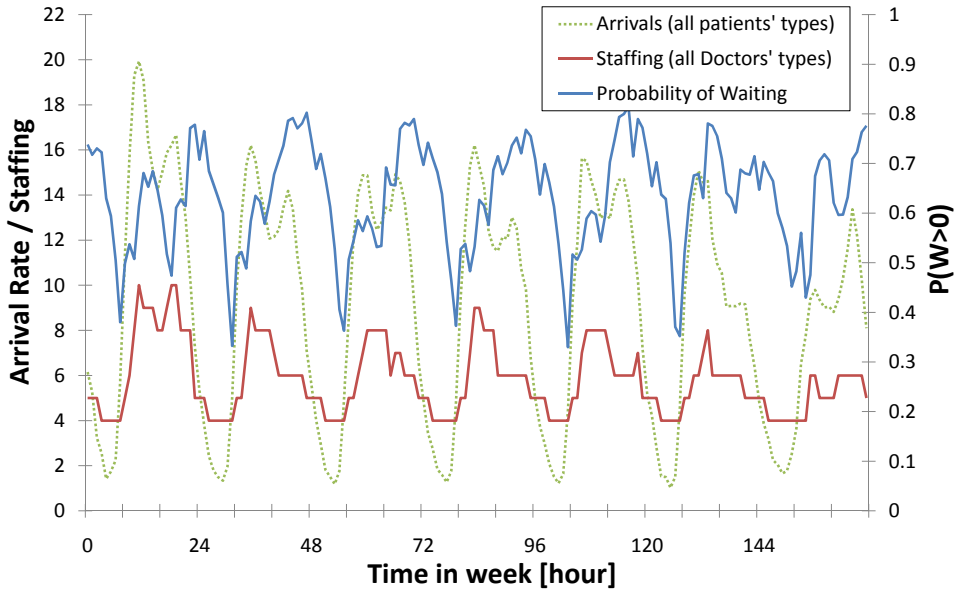


Figure 23: EW case study - Arrivals, staffing, and waiting when $\beta = 0.1$

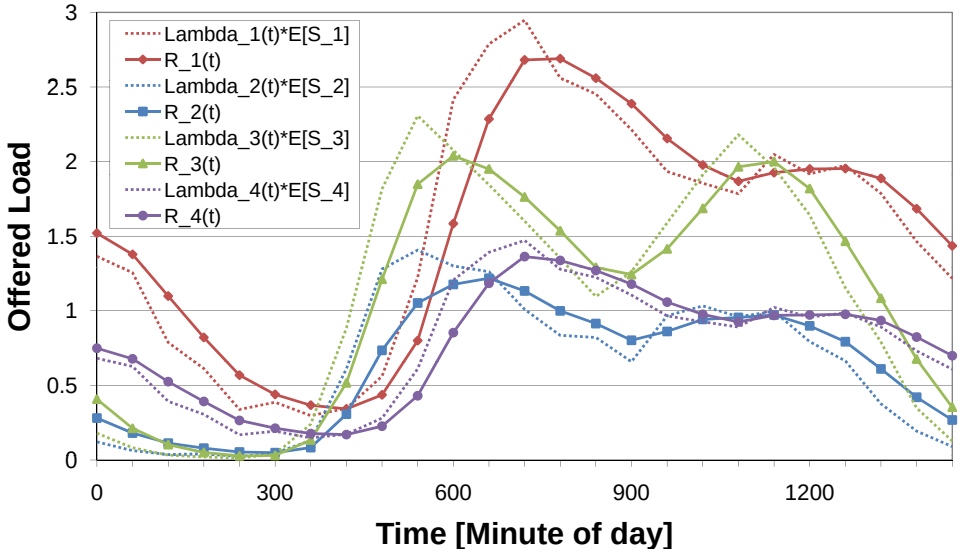


Figure 24: EW case study - Offered load vs. RCCP

8 Approximating the Number of Needy Customers and Waiting Times in the QED Regime

In this section, we derive approximations for the actual number of customers in the system and the virtual waiting time process. Usually, one uses fluid and diffusion approximations for this purpose. In appendix B we develop the fluid and diffusion approximations for our system, based on the mathematical framework of Mandelbaum et al. [53] on time-varying queues, and show their usefulness for analyzing mass-casualty events. Note that while it is clear that fluid approximations are very useful in analyzing time-varying systems, these approximations are also useful in understanding the transient behavior of systems with time-constant parameters [54]. For example, we might need to evaluate the probability that the number of customers (patients) in the system will exceed a certain threshold during a certain time horizon. This is useful when setting control rules for the EW, for example starting special procedures such as ambulance diversion and calling for additional staffing. The answer to such questions requires diffusion approximations, such as the ones we develop in Appendix B.

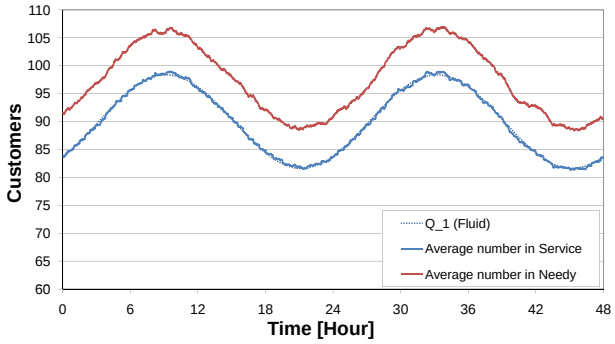
These fluid and diffusion approximations are known to work well under the zero-measure assumption (i.e. under the assumption that the time the system spends in critically-loaded values is negligible; see for example Mandelbaum et al. [55]). In our case, when the system operates in the QED regime, the system is critical at **all** times, and furthermore, the accuracy of QED approximation was not examined. The problem when using these approximations in the QED regime is twofold: first, we have numerical difficulties in calculating the diffusion process itself since the diffusion approximation is none-autonomous. Second, the fluid process itself has a different interpretation under the QED regime: no longer does it represent the average behavior of its originating stochastic system.

To understand the interpretation problem, we use the following example from Case Study 1. The left diagram in Figure 25 shows the fluid solution of the process $Q_1^{(0)}(t)$ (the number of Needy customers), as well as the following simulation results: the average number of customers in the Needy state, and the average number of customers in service. We see that, although the fluid model is supposed to represent the number of Needy customers in the QED regime, it fits perfectly the number of customers *in service* and ignores the number of customers waiting in queue (for service). This is because our MOL staffing procedure keeps the staffing level always slightly above the average number of customers. Thus, the fluid approximation sees the system as if it had an infinite number of servers, and actually calculates the number of busy servers, without the queue.

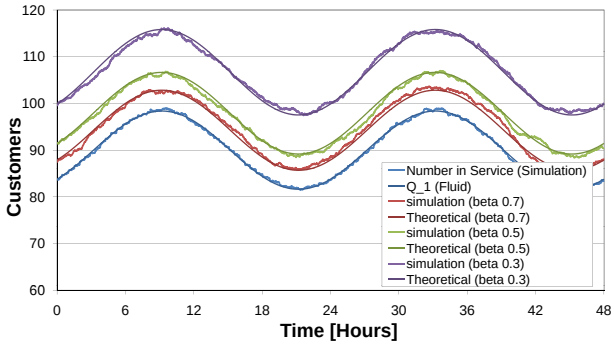
In order to fill the gap and to estimate correctly the number of Needy customers (in queue and in service), we will use the insight we obtained previously, namely, that under MOL staffing, the system behaves as if the Needy state were a stationary M/M/s model (Erlang-C). Therefore, we can use the stationary approximation of the Erlang-C model to estimate the number of customers in the queue. Halfin and Whitt [39] approximated $E[Q(\infty)]$ by the following formula: $E[Q_1(\infty)] = \frac{\lambda}{\mu} + \alpha\frac{\lambda}{s\mu}\left(1-\frac{\lambda}{s\mu}\right)^{-1}$ where $\alpha = P(W>0) = \left[1+\beta\frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}$. We propose an MOL correction, to adjust this formula to time-varying environments in the following way:

$$E[Q_1(t)] = R(t) + \alpha\frac{R(t)}{s(t)}\left(1-\frac{R(t)}{s(t)}\right)^{-1}.$$

The right diagram in Figure 25 compares the corrected approximation to simulation results for various β values. We see that the simulation and approximation are remarkably close.



Fluid vs. Simulation



Corrected approximation vs. simulation

Figure 25: $Q_1(t)$ - Fluid approximation vs. simulation results under QED staffing, for various β's

We can also provide a correction to the E[W] function in the QED regime, using the following expression:

$$E[W(t)] = \frac{\alpha}{\mu s(t)}\left(1-\frac{R(t)}{s(t)}\right)^{-1}.$$

Experiments show that this correction works well for $\beta > 0.3$, as can be seen in Figure 26.

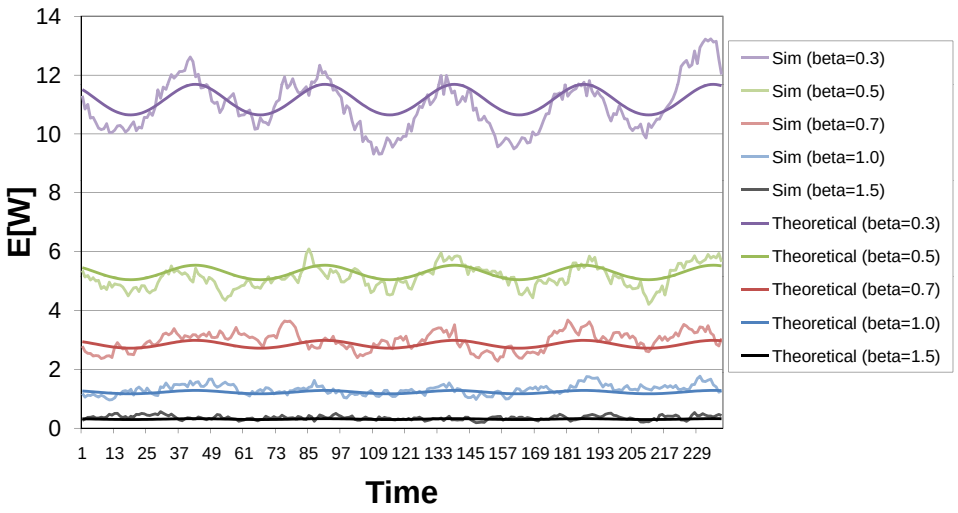


Figure 26: $E[W(t)]$ - Corrected fluid approximation vs. simulation for various β's

9 Erlang-R: Conclusions and Future Research

In this first part, we developed a model that incorporates discontinuous services in time-varying environments. We proposed an appropriate staffing procedure that balances the system in the QED regime in a way that stabilizes both service-level performance measures and utility over time. We validated our approach using simulation, and showed that it works well both in large and small systems. We also showed that it works well in realistic cases, using an Emergency Ward simulation. The Erlang-R model is general enough to capture various EW setting such as fast-track, triage etc. In fact, we also incorporated natural operational constraints, such as minimal staffing levels and maximal frequency of staffing changes. The problem of incorporating additional constraints, for example those that result from regulations, is left for future research. This will require the identification of constraints that can be accommodated by MOL-based methods, and then applying some method of optimization via simulation [27, 7].

We demonstrated, both analytically and by examples, that using a simpler model (Erlang-C) in Erlang-R situations could be detrimental. We also identified the circumstances in which the differences are more significant. Lastly, we developed Taylor-series approximations to the offered-load measure, for cases where it cannot be calculated explicitly.

In the future, we plan to simulate additional case studies, which will validate our model in more complex situations, such as deterministic service times and log-normal service times.

We are also interested in the effect of priority schemes on staffing since, in some cases, one may wish to give priority to patients that are on their first visit, or, alternatively, to patients who have been in the system for a long time. Other priority schemes could (or rather must) take into account clinical information, for example, the severity of the medical condition. More investigation is also needed to find exact formulae for the relation between β and $P(W > 0)$ for small systems.

An important extension could be made to accommodate abandonment of customers into our model. This is a much more complex situation than in the simple Erlang-C model, since abandonment can take different patterns. In some cases, the customer will abandon only in the first waiting; this is known as the Left Without Being Seen (LWBS) effect. In the hospital with which we worked, we observed a different abandonment phenomenon: Some patients abandon the EW somewhere *between* services. No one knows exactly in what stage, since these patients took their medical records with them, and no one knows exactly why: was it due to the lengthy waiting period, dissatisfaction with the medical staff, or financial reasons. We cannot simply assume, therefore, that patients decide to abandon only due to waiting, and that abandonment rates are equal for the first

and the returning visits. Lastly, we are also planning to analytically investigate the convergence of the diffusion process under our staffing procedure. This may shed some light on the reasons why our time-varying staffing rule works so well.

In subsequent parts, we extend the Erlang-R model by setting an upper limit on the number of customers within the system. This situation is more complex and we model it via a semi-open queueing network.

Part II

The Semi-Open Erlang-R Model

10 Introduction

10.1 QED Queues in Internal Ward Application

In this part of the thesis, we examine a different network with ReEntrant customers. It was build to model Internal Wards (IW), in which the number of patients in the system is bounded. One can consider this model also as an extension of the Nurse-To-Patient model of Jennings and de Véricourt [45]. This extension accommodates jointly bed allocation and nurse staffing, in the QED regime. Bed allocation determines blocking probabilities of the MUs; nurse staffing determines the delay probabilities of patients waiting for medical care inside the MUs. The combination of these two issues will allow us to determine the appropriate capacity planning while gaining a deeper understanding of the relationship between bed allocation and nurse staffing, which are usually considered separately.

More explicitly, we consider the following medical unit: the maximal number of beds available is n and the number of nurses serving patients in the unit is s. Typically, the number of nurses is fewer than the number of beds, i.e. $s \leq n$, which is assumed here as well. Patients in the unit require the assistance of a nurse from time to time. When such assistance is required we refer to the patient's state as *Needy*. Otherwise, we call the patient's state *Content*. When patients arrive, they start in a needy state and then alternate between needy and content states. When patients are discharged from the hospital, they leave from the needy state. The last treatment can thus reflect the discharge process. After a patient leaves that medical unit, his bed needs to be made available for a new patient. This is usually done by a cleaning crew and not by the nurses of the unit. We will refer to that state of a bed as *Cleaning*. When patients become needy and an idle nurse is available, they are immediately treated by a nurse. Otherwise, patients wait for an available nurse. The queueing policy is FCFS. After completing treatment, a patient is discharged from the hospital with probability $1 - p$ or goes back to a content state with probability p until additional care procedures are required. We assume that the treatment times, are independent and identically distributed (i.i.d.) as an exponential random variable with rate μ and that the content times are also i.i.d. exponential with rate δ. As noted above, after the discharge process the bedding must be changed. We assume that cleaning times are i.i.d. exponential with rate σ. We also assume that

the needy, content, and cleaning times are independent of each other and of the arrival process.

The arrival process is assumed to be a Poisson process at rate λ, constant over time. Another major assumption concerning the arrivals is that if patients arrive at the MU in order to be hospitalized but the unit is full, they are diverted elsewhere, for example back to the Emergency Ward (EW) or to other units of the hospital. Thus, one can view this as blocking of the unit; in this situation we say that the MU is *blocked* and the request is *lost* (In a call center such a situation corresponds to customers encountering a busy signal).

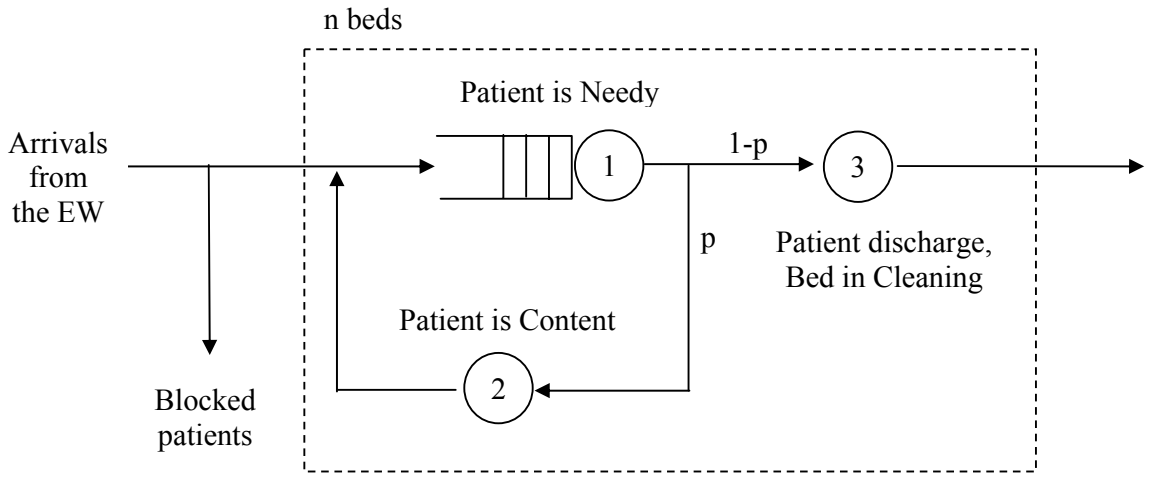


Figure 27: The IW model as a semi-open queueing network

Remark: We would like to point out that the parameter n need not represent a constraint on the physical capacity of an MU. Rather, it could also stand for an operational constraint on the number of patients that can reside simultaneously within the system (in analogy to CONWIP [70, 4]). Specifically, such a constraint would be reasonable in out of the EW, heavily-loaded EW's, where walking-patients can be delayed, if there is no risk for their deterioration. We model the MU as a semi-open queuing system (see Figure 27) in Section 11 we reduce it to a closed Jackson network, as will be shown later (see Figure 31), which yields a product-form steady-state. In Chapter 12 we define some system measures to this network. These system measures are designed to enable us to answer the following questions:

1. How many nurses should be planned for the unit? One can ask this in the context of either providing reasonable service levels, or from the viewpoint of cost / profit optimization. An answer to this question can be based, for example, on the following measures:

   (a) What is the probability of waiting for a nurse?

(b) What is the probability to wait more than T units of time?

2. How many beds should be planned for in the unit? This question could also be answered from the viewpoint of service quality, i.e., providing reasonable availability, or from the viewpoint of cost optimization. Again some measures should be calculated such as:

   (a) What is the probability of blocking? i.e., the fraction of time that the system is in full capacity, which translates into the percentage of patients not admitted upon arrival.

   One should notice that the blocked patients are getting stuck in the EW which, in turn, can result in reducing the available capacity of the EW itself, as well as hurting patients' safety and well-being.

Naturally, one would like the answers for Questions 1 and 2 to be synchronized.

Next, in Chapter 13, we define QED scaling for the system in Figure 27 in the following way:

$$s = \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}} + o(\sqrt{\lambda}), \qquad -\infty < \beta < \infty \qquad \text{(i)}$$

$$n - s = \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} + o(\sqrt{\lambda}) \qquad -\infty < \eta < \infty \qquad \text{(ii)}$$

where p,μ,δ, and γ are fixed model parameters. Term (ii) corresponds to requests queueing for a nurse, and Term (i) corresponds to the effective capacity in the non-queue states. Now, the QED limits of our performance measures can be calculated. For example, as λ, s and n increase indefinitely and simultaneously, according to the above QED scalings, and $\beta \neq 0$, then

$$\lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}\right)^{-1},$$

where W denotes waiting for nurse-service, $\eta_1 = \eta - \beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}$; $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively. This limit, together with several other QED limits, are proven in Section 14. The result supports our definition of the QED regime (non-degenerate delay probability). These limits enables us to better understand the dynamics of the system and derive managerial insights of the behavior of the system in the QED regime.

Our model is closely related to the one in Khudyakov [49], where it was developed for a call center with an Interactive Voice Response (IVR) system. In fact, all our heavy-traffic approximations have the same structure as those in [49]. With this observation as a starting point, we introduce in Section 17 a generalization that covers both systems, as well as some additional modeling possibilities of

our MU system. Note that Khudyakov's model is general, in the sense that it covers various other models such as the M/M/S/N loss system, M/M/N/N (Erlang-B), and M/M/S (Erlang-C). Our generalization covers any semi-open network, with n spaces, with one service station with s servers, and any finite number of delay procedures. We use it to develop steady-state approximations of the time-varying semi-open Erlang-R model.

10.2 The Time-Varying Semi-Open Erlang-R Model

The IW model and the Erlang-R model are closely related. We can consider the IW model as a generalization of the Erlang-R model with an additional upper bound on the number of customers in the system (without the cleaning state). Therefore, we call it *semi-open* Erlang-R. Figure 28 depicts a graphical representation of this system. This model can also be used to represent EW in which there is a restriction on the number of beds. In that case, the number of customer (patients) is bounded by the number of spaces (beds) in the EW, which is n. Customers that are blocked are thought of as being transfered to another system, for example using ambulance diversion procedures.

We wish to understand the significance of Reentrant customers in semi-open systems, as we did for the open system. In this case, the natural comparison is not to the Erlang-C (M/M/s) model, but to a corresponding loss system (M/M/s/n). Using our generalization from Section 17, we show in Section 19 that the steady-state distribution of the semi-open Erlang-R system (Needy state) and a loss system are different. The two systems become similar as the offered load ratio $\left(B = \frac{\delta}{p\mu}\right)$ increases. We then develop MOL staffing rule for the time-varying semi-open Erlang-R model. It is based on the offered load of the open Erlang-R model and the steady-state QED approximations of the semi-open Erlang-R model. Using simulation we demonstrate that this approach works very well in the QED regime, and stabilize both $P(W > 0)$ and P(block) over time.

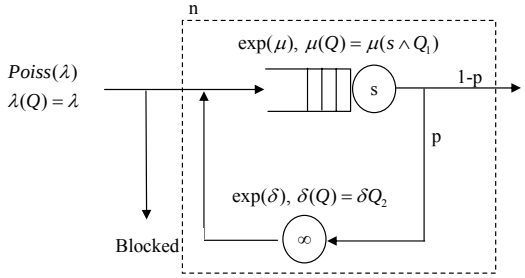


Figure 28: The semi-open Erlang-R model

10.3 Literature Review

10.3.1 Background on System Design

Due to the complexity of the Health-Care system, capacity management decisions are carried out hierarchically. We will now shortly describe the whole process: forecasting demand, setting bed capacity at the hospital level, setting the allocation of beds inside each individual hospital, setting staffing levels, shifts scheduling and rescheduling. It is common practice to distinguish between static- and dynamic-capacity decisions; usually these decisions are the charter of different management teams. While static-capacity is hard to change and planning is made for long-term periods, dynamic-capacity is flexible: namely, it can be adapted to changes in circumstances within a short period of time. In the literature overview on capacity planning in health care, presented by Smith-Daniels et al. [68], the following classification is proposed: facility resources planning (for example: bed allocation) is separated from work-force resource planning (such as nurses and doctors staffing and scheduling). In our literature review we use the same classification.

10.3.2 Managing Bed Capacity

Long-term capacity planning is based on forecasting the demand for inpatient services. The forecast is based on mathematical models (such as time series) that predict the changes in inpatient demand over long periods (i.e. months and years). For example, one can use the forecasting models of Jones et al. [47], or Kao and Tung [48]. Based on this prediction policy makers can deduce the required bed capacity of the hospital.

As in other service systems, in most Health-Care systems, the arrival rate of patients entering the system varies over time. Over short periods of time, minute-by-minute for example, there is significant *stochastic* variability in the number of arriving patients. Over longer periods of time, the course of the day, the days of the week, the months of the year – there can also be *predictable* variability, such as the seasonal patterns that arriving patients follow. Example of patterns in admission of cardiac inpatients into EW, during an average day, can be found in de Bruin et al. [19]. Harper and Shahani [42] have shown another example of changes in mean bed occupancy through the months of the year of adult medical (as opposed to surgical) population, in a major UK NHS Trust; this pattern could reflect the pattern of admissions, assuming that the bed capacity was fixed during that period.

Because the service capacity cannot be inventoried, one should vary the number of available beds and medical staff in the short-term, to track the predictable variations in the arrival rate of patients.

If we do that, we are able to meet demand for service at a low cost, yet with acceptable delay times and acceptable blocking rates. But despite of these patterns, it was common practice in the US, for many years, to determine hospital bed capacity by using the mean bed occupancy measure. This was done both by policy-makers and various levels of hospital management. Green [31] showed by simple M/M/S queueing model that this method was wrong; in Europe, Harper and Shahani [42] claiming the same, developed a simulation tool for fitting acceptable bed occupancy to the monthly and daily arrival-rate patterns; the tool is based on the Length-of-Stay (LOS) statistics, and the refusal rates. In Israel, bed capacities are determined by turnover rates per bed, and the forecasting of future mean LOS.

The applications of Queueing model to solve beds' allocation problems in not new. Green [31] suggested the use of a simple M/M/S model while de Bruin et al. [18] proposed to solve the beds allocation problem via a loss system (Erlang-B M/M/n/n). The latter is due to the fact that the number of beds in the ward is limited, and if there is no space for a patient in the ward, the patient is transformed to an alternative location for hospitalization.

As explained, the long- and short-term analysis of beds requirements are helping to determine the appropriate bed allocation. Hospitals distinguish between maximal bed capacity and nominal bed capacity. The former is the true physical constraint of the system, while the system is designed to operate with the latter. Naturally, the maximal bed capacity itself is mostly fixed (in scale of months). Therefore, the available capacity of the MU is practically determined by more flexible elements such as the number of doctors and nurses.

10.3.3 Managing Work-Force Capacity

The work-force of a hospital includes nurses, doctors, laboratory workers and others. Most of these human resources need very long and expensive training, and together contribute as much as 70% of the hospital expenses. Nursing salaries make up the largest single element in hospital costs [67]. Thus, much attention is needed in managing the work-force capacity.

At the top of the work-force planning hierarchy, a long-term staffing problem is solved to ensure that monthly staffing requirements are met. The problem is usually considered at the management level, considering costs and the annual rate of personnel turnover, which reflects dissatisfaction, differences in workload between wards, and seasonal variation in admission rates. Hospital staffing involves determining the number of personnel of the required skills in order to meet predicted requirements. It is sometimes referred to as nurse budgeting, or workforce scheduling in other personnel planning environments. Burke et al. [13] reviewed some of the work on this subject.

One can determine, using queueing models, the number of nurses which should be available to serve patients over a given time slot. The staffing levels can vary between shifts or months, and track the predictable variations in the arrival rates of patients. But so far, much more robust structures are used; in 2004 in the US, the California Department of Health Services (CDHS) published a law that specifies a nurse-to-patient ratios that determine the minimal staffing levels allowed [63]. In other countries, such as Israel, staffing levels are determined by labor agreements. We will specify only the last development in the field of Health-Care staffing models; in 2006, Jennings and de Véricourt [45] used a queueing model, to develop new nurse-to-patient ratios, that are a function of the MU size. These ratios were developed in the QED regime in order to balance the work-force efficiency and the quality of care.

The next stage is to determine each nurse's shifts using scheduling models. This planning stage is often referred to in the literature as the Nurse Rostering Problem (NRP) or the Nurse Scheduling Problem (NSP). Cheang et al. [16] defined the NRP as a procedure which involves producing a periodic (weekly, fortnightly, or monthly) duty roster for nursing staff. The schedules are often restricted by legal regulations, personnel policies, nurses' preferences and many other requirements that may be hospital-specific. These can be quite complex. Naturally, one of these constraints is the minimal staffing level needed to satisfy the service standards, calculated in the previous stage. There are several reviews of the different methods for NRP, the most recent being those of Cheang et al. [16] and Burke et al. [13]. There are also some general survey papers in the area of personnel rostering such as that of Ernst et al. [24].

After the scheduling phase comes the third step, which represents the lowest level of the hierarchy: the reallocation of nurses. This phase is a fine-tuning of staffing and scheduling. It involves determining how float nurses are assigned to units based on nonforecastable changes or absenteeism. See, for example, Bard and Purnomo [8].

10.4 QED Queues in Internal Wards

As opposed to the hierarchy noted above, we suggest a unified method that will determine the bed allocation and nurse staffing levels simultaneously, in the QED regime. We shall focus on QED queues in order to balance patients clinical needs for timely service with the economical preferences of the system to operate at maximal efficiency. We have shortly described the QED regime, and its relevance to our environment, in Section 1.

The only work that viewed hospital queues in the QED regime is that of Jennings and de Véricourt [45]. They analyzed the prevalent staffing practice of an *a priori*-fixed patient-to-nurse

ratio (for example, 6-children-per-nurse in a pediatric ward). They showed that such a practice results in either over- or under-staffing in large or small MUs respectively, which can be remedied by square-root staffing. Their mathematical framework [46] is a special Jackson closed-network (machine-repairmen) model of the MU, where the circulating customers are patients' requests for nursing assistance. (Randhawa and Kumar [66] is a related model, where losses replace the delays in [45, 46].)

Jennings and de Véricourt [45, 46] considered the MU model as depicted in Figure 29. It is, in fact, a $M/M/s/\infty/n$ queueing model. Specifically, there are n beds, all occupied by patients. From time to time, these patients require the assistance of one of s nurses, in which case we refer to the patients' state as *needy*. Otherwise, their state is *content*. The state of patients alternate between needy and content states. When patients become needy and an idle nurse is available, they are immediately treated by a nurse. Otherwise, patients wait for an available nurse. The queueing policy is FCFS. It is assumed that treatment times, i.e. needy-state times, are i.i.d. exponential with rate μ; content times are also i.i.d. exponential with rate λ. It is also assumed that the needy and content times are independent of each other.

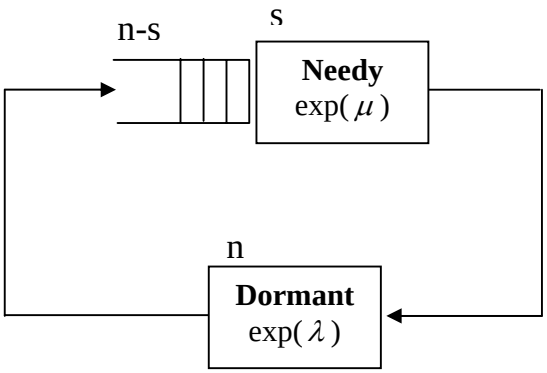


Figure 29: Jennings and de Véricourt's model [45, 46]

Jennings and de Véricourt define the operation regimes for that system as follows: Let us define s_n as the number of nurses in the n-th system, $\bar{s} = \lim_{n\to\infty} \frac{s_n}{n}$, and $r = \frac{\lambda}{\lambda+\mu}$ then

- If $\bar{s} < r$, the system operates in an Efficiency Driven (ED) staffing regime ($T > 0$)
- If $\bar{s} = r$, the system operates in a QED staffing regime ($T \geq 0$ and small)
- If $\bar{s} > r$, the system operates in a Quality Driven (QD) staffing regime ($T = 0$)

where T is a fixed parameter, representing the required time for service. Then the appropriate

QED staffing rule is: $s_n = \lceil rn + \beta\sqrt{n} \rceil$. Naturally, the QED limit of the probability of delay was calculated, and a central result of their article [46] is their

Proposition 3. *The approximate probability of delay has a nondegenerate limit* $\alpha \in (0,1)$ *if and only if* $\beta_n = \left(\frac{s_n}{n} - r\right)\sqrt{n} \to \beta$, *as* $n \to \infty$, *for some* $\beta \in (-\infty, \infty)$, *with*

$$\alpha = \left(1 + e^{\frac{-\beta^2}{r^2}} \sqrt{r} \frac{\Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right)}{\Phi\left(\frac{-\beta}{r\sqrt{\bar{r}}}\right)}\right)^{-1}.$$

Here $\bar{r} := 1 - r$.

Significantly, the QED regime is *many-server asymptotic*, as the number of servers increases indefinitely. Yet it is also relevant for application in small systems, including nurse staffing, being able to accommodate a small number of nurses (*single-digit* and above). This relevance is a consequence of the surprising accuracy of square-root staffing, a fact discovered in [11] and also recently supported by the results of Jennings and de Véricourt [45].

As mentioned before, the QED regime arises naturally as the mathematical framework for patient-flows from the EW to the MUs. Indeed, consider the queueing times at the EW (resulting in MU hospitalization) vs. the consequent Length-of-Stay (LOS) at the MUs: hours vs. days is typical. Also, the number of beds (servers) in MUs of moderate-to-large hospitals is in the 10's (35-50 beds in each of 5 MUs, at the Technion affiliated hospital) - which is well within the accuracy limits of QED asymptotics.

10.5 Research Objectives

The remainder of this part is organized as follows: the Internal Ward model we developed is introduced in Section 11. System measures are defined in Section 12. The QED regime of the system is described in Section 13. The development of heavy traffic limits, in the QED regime, of our system-measures are detailed in Section 14. In Section 15 we compare the approximations with the exact calculations, trying to define the ranges where the approximation is most accurate. In Section 16 we compare our model with other queueing models investigated in the past. A generalization of our model and the appropriate QED asymptotics are presented in Section 17. A method for defining Optimal Design is shown in Section 18. We then investigate in Section 19 the semi-open Erlang-R model in time-varying environments, and compare it to a loss system. Section 20 conclude the analysis with some managerial insights of the behavior of the system in the QED regime. Finally, conclusions and suggestions for further research are discussed in Section 21.

11 An Extended Nurse-to-Patient Model

11.1 The Medical Unit: Internal Ward (IW)

We consider the following medical unit: the maximal number of beds available is n and the number of nurses serving patients in the unit is s. Typically, the number of nurses is fewer than the number of beds, i.e. $s \leq n$, which is assumed here as well. Patients in the unit require the assistance of a nurse from time to time. When such assistance is required we refer to the patient's state as *needy*. Otherwise, we call the patient's state *dormant*. When patients arrive, they start in a needy state and then alternate between needy and dormant states. When patients are discharged from the hospital, they leave from the needy state. The last treatment can thus reflect the discharge process. After a patient leaves that medical unit, his bed needs to be made available for a new patient. This is usually done by a cleaning crew and not by the nurses of the unit. We will refer to that state of a bed as *cleaning*. When patients become needy and an idle nurse is available, they are immediately treated by a nurse. Otherwise, patients wait for an available nurse. The queueing policy is FCFS. After completing treatment, a patient is discharged from the hospital with probability $1 - p$ or goes back to a dormant state with probability p until additional care procedures are required. We assume that the treatment times, are i.i.d. as an exponential random variable with rate μ and that the dormant times are also i.i.d. exponential with rate δ. As noted above, after the discharge process the bedding must be changed. We assume that cleaning times are i.i.d. exponential with rate σ. We also assume that the needy, dormant, and cleaning times are independent of each other and of the arrival process.

The arrival process is assumed to be a Poisson process at rate λ. Another major assumption concerning the arrivals is that if patients arrive at the MU in order to be hospitalized but the unit is full, they are diverted elsewhere, for example back to the EW or to other units of the hospital. Thus, one can view this as blocking of the unit; in this situation we say that the MU is *blocked* and the request is *lost*. (In a call center such a situation corresponds to customers encountering a busy signal).

To this end, the MU is modeled as the *semi-open* queueing network. For reading convenience, we reconstruct the IW model form Figure 27, here as well.

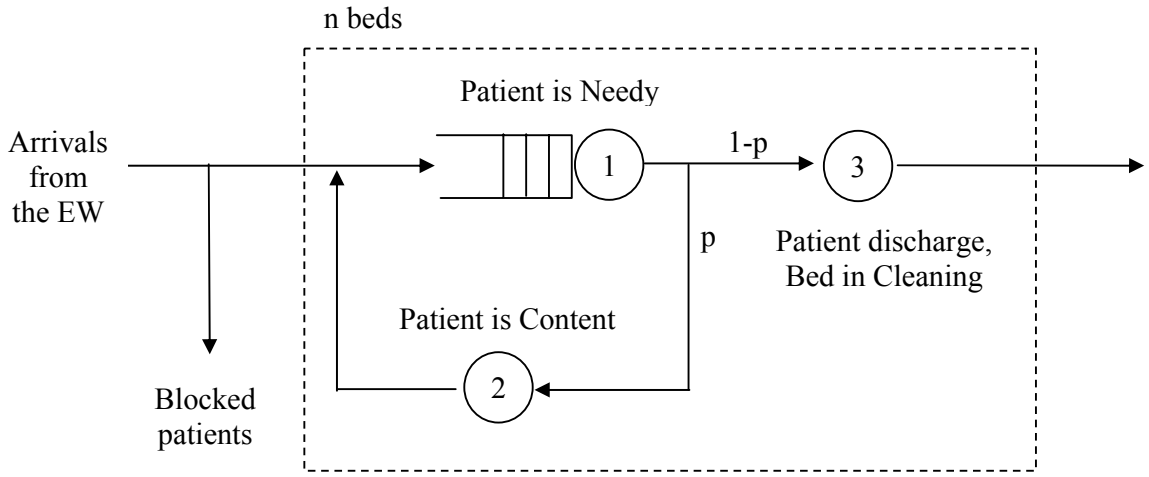


Figure 30: The IW model as a semi-open queueing network

11.2 The Model

The analysis of the above semi-open network can be reduced to that of a closed Jackson network[1], which yields a product-form steady-state. A scheme of our model in this form appears in Figure 31. This is done by representing our model of the medical unit as a system with four nodes. Node 1 represents beds with patients in a *needy* state. Node 2 represents beds with patients in a *dormant* state. For convenience, we sometimes refer to a bed with a patient as simply a patient. Node 3 represents beds in preparation, i.e. in a *cleaning* state. Node 4 represents prepared beds, awaiting a patient. Nodes 1 to 3 are all multi-server queues. The first node can handle, at most, s patients at one time. The second and third nodes can "handle" or contain at most n patients at a time. Node 4, which is a single-server queue, represents the external arrival process into the unit as will be explained later.

Let $Q(t) = (N(t), D(t), C(t))$ represent the number of beds in the *needy*, *dormant* or *cleaning* states respectively. Since n is the maximum number of patients/beds in the system, i.e. needy,

[1]A Jackson network consists of several interconnected queues; it contains an arbitrary but finite number N of service centers, each has an infinite queue. Let i and j denote service stations in that network. The service discipline is FCFS, where the service time in station i is drawn independently from the distribution $exp(\mu_i)$; (we could have state dependent service rates $\mu_i(n_i)$). Customers travel through the network according to transition probabilities. Thus, a customer departing from station i chooses the queue in station j next with probability P_{ij}. All the customers are identical; they all follow the same rules of behavior. If the network is open then the arrivals from outside to the network (source) arrive as a Poisson stream with rate λ, and from each node there is at least one path to exit, i.e. the probability that a customer entering the network will ultimately depart from the network is 1. If the network is closed, there are no arrivals or departures hence, there is a constant population of customers in the network.

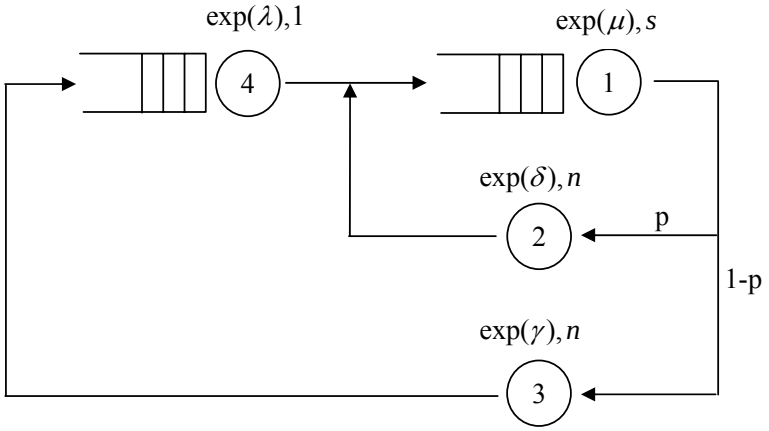


Figure 31: The IW model as a closed Jackson network

dormant or in cleaning, then $N(t) + D(t) + C(t) \leq n$, for all $t \geq 0$. The process Q is a finite-state continuous-time Markov chain. The states of the chain will be denoted by the triplets $\{(i, j, k)|i + j + k \leq n, i, j, k \geq 0\}$. A state (i, j, k) represents a situation where i *needy* patients are being served or wait for service, j *dormant* patients are in the unit but need no service at the time, and k beds are being prepared for future patients. $i + j + k \leq n$.

Our medical unit model can be viewed as part of a closed Jackson network. The first node (*needy*) can be modeled as s servers with a queue in front of them, with an exponential service time at a rate of μ. The second node (*dormant*) is an infinite-server node, with an exponential service time at a rate of δ. The third node (*cleaning*) is also an infinite-server node, with an exponential service time at rate of γ. A new patient can enter the unit only if $N(t) + D(t) + C(t) < n$. Thus, the process of admitting new patients into the unit has the intensity:

$$\lambda(i, j, k) = \begin{cases} \lambda & if \;\; i + j + k < n, \\ 0 & \text{otherwise.} \end{cases}$$

In order to formulate this situation a fourth node has been added in which we have a single exponential server of rate λ.

Generally, this type of a closed Jackson network has the following product form solution for its stationary distribution [30]:

$$\pi^0(i, j, k, l) = \begin{cases} \frac{\pi^1(i)\pi^2(j)\pi^3(k)\pi^4(l)}{\sum_{a+b+c+d=n} \pi^1(a)\pi^2(b)\pi^3(c)\pi^4(d)} & , \quad i + j + k + l = n, \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Here $\pi^m(i)$ is the steady state probability for node m, $m = 1, 2, 3, 4$ (M/M/s, M/M/∞, M/M/∞, M/M/1 respectively). The stationary probability $\pi(i, j, k)$ of having i *needy* patients, j *dormant*

patients and k beds in *cleaning* can thus be written in a product form as follows:

$$\pi(i,j,k) = \begin{cases} \pi_0 \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k & , \quad 0 \le i+j+k \le n, \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Here $\nu(i)$ is defined as

$$\nu(i) := \begin{cases} i! & , \quad i \le s, \\ s!s^{i-s} & , \quad i \ge s, \end{cases}$$

where π_0 is given by (see Appendix C.1)

$$\begin{aligned} \pi_0^{-1} &= \sum_{0 \le i+j+k \le n} \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k \\ &= \sum_{l=0}^{n} \frac{1}{l!} \left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l \\ &+ \sum_{l=s+1}^{n} \sum_{m=s+1}^{l} \sum_{i=s+1}^{m} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \frac{1}{(m-i)!(l-m)!} \left(\frac{\lambda}{(1-p)\mu}\right)^i \cdot \\ &\qquad \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \left(\frac{\lambda}{\gamma}\right)^{l-m}. \end{aligned} \tag{11.1}$$

Note that π is also a function of n and s. In order to emphasize this dependence, we shall sometimes use $\pi_n(\cdot), \pi_{n,s}(\cdot)$, etc.

In this work we would like to focus on some managerial questions such as:

1. How many nurses should be planned for the unit? One can ask this in the context of providing reasonable service levels, or from the viewpoint of cost / profit optimization. An answer to this question can be based, for example, on the following measures:
   (a) What is the probability of waiting for a nurse?
   (b) What is the probability of waiting more than T units of time?
2. How many beds should be planned for in the unit? This question could also be answered from the viewpoint of service quality, i.e., providing reasonable availability, or from the aspect of cost optimization. Again some measures should be calculated such as:
   (a) What is the probability of blocking? i.e., the amount of time when the system is in full capacity ($i+j+k=n$), which translates into the percentage of patients not admitted into the MU.

Naturally, one would like the answers for Questions 1 and 2 to be synchronized.

11.3 Alternative Models

In the following chapters we will conduct a stationary analysis of the above stated model. Nevertheless, we also suggest alternative ways to model the system, which we will use later for our non-stationary analysis. In this chapter we defined several additional possibilities to model the system. These models are illustrated by figures that show only the dormant and needy states (i.e. without cleaning). In the following subsections, Q_i will always represent the number of patients at node i.

11.3.1 Proposal 1

In this first proposal, we assume that the arrival rate into the IW is linearly related to the occupancy level of the ward; if the occupancy increases, the arrival rate decreases. This assumption capture various effects: First, when total load is normal, if the hospital operates in a parallel setting, (i.e. there are a few identical IWs in the hospital), there is usually someone that balances the system by transferring patients among wards. Second, when the total load is high, there are balancing effects that reduce arrival rates, such as diversions to other hospitals, and doctors that refrain from referring additional patient during over-loaded periods.

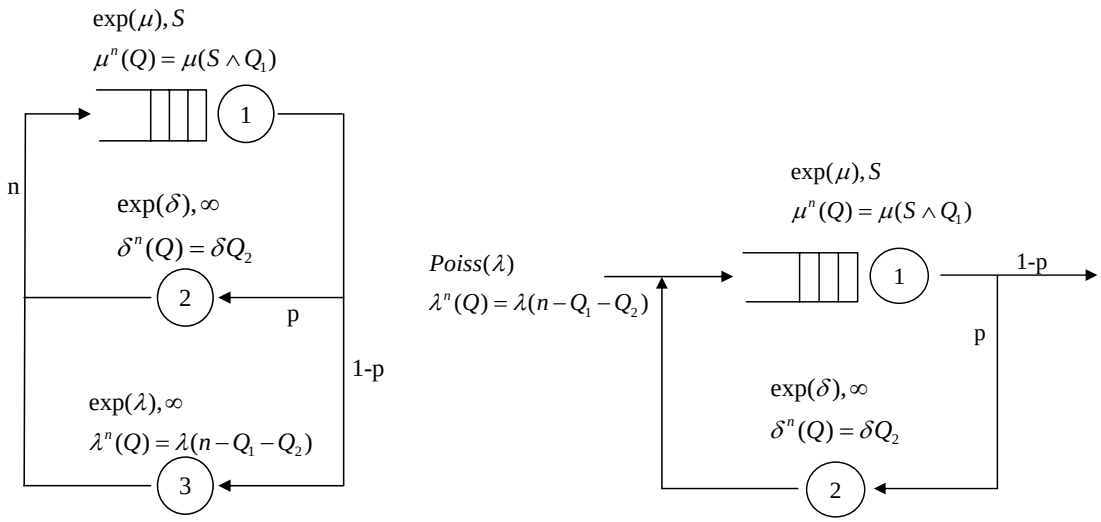
The system is presented in Figure 32 in two alternative versions. We regard the situation as a closed network, with a state-dependent arrival rate, in which $\lambda(Q) = \lambda \cdot (n - Q_1 - Q_2)$, where Q_1 is the number of patients in the needy state, Q_2 is the number of dormant patients, and n is the number of beds in the ward.

11.3.2 Proposal 2

In this possibility, the arrival rate is fixed as long as there is a bed available in the system. This is equivalent to the model presented in section 11.2, but without the cleaning state. The system is presented in Figure 33.

11.3.3 Proposal 3

We propose another model that is presented in Figure 34. Here we relax the constraint on n and ask the following question: What should s and n be so that the probability of waiting is less than α and the probability of exceeding n is less than β, where the interpretation of having more than n beds is that patients are attended to in the hospital corridors, for example. This is a realistic scenario, since in spite of the fact that bed-allocation is considered as static-capacity, there is some



Version 1

Closed Jackson Network

Version 2

Figure 32: Alternative model - Proposal 1

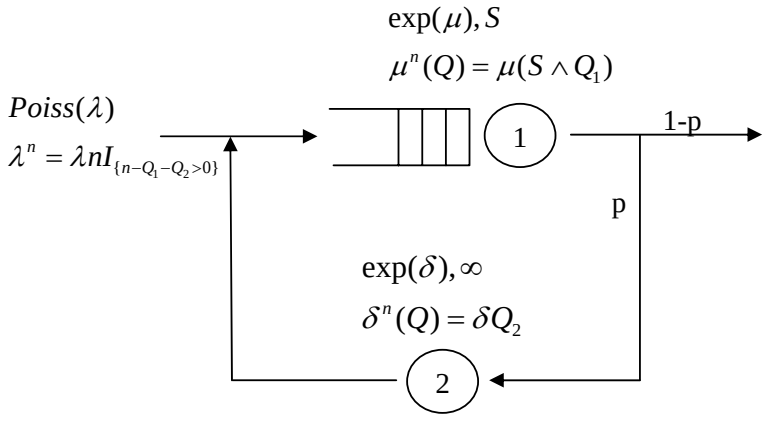


Figure 33: Alternative model - Proposal 2

flexibility in managing this resource; in time of need, one can add beds in rooms and corridors. In addition, this alternative might be easier to solve. In this way μ is state dependent but λ is not, and the system is purely open.

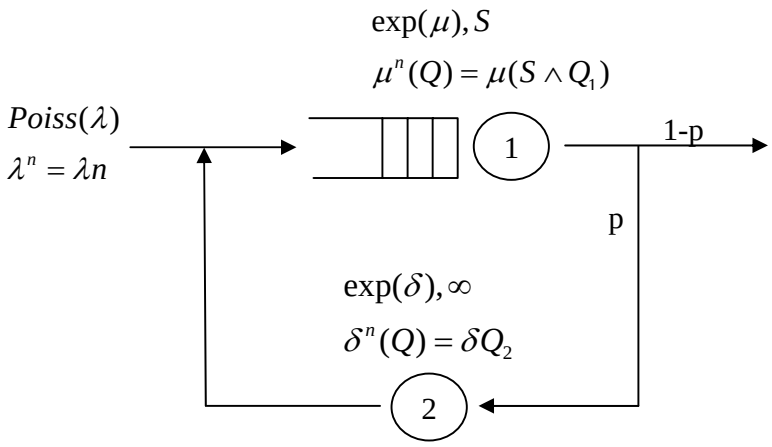


Figure 34: Alternative model - Proposal 3

11.3.4 Proposal 4

In this model, which is presented in Figure 35, we separated the LOS of patients from the service inside the medical department. The LOS is assumed to be exponential with mean $1/\nu$ and the size of the population is restricted to n. When the patient is in the system he alternates between the dormant and the needy states.

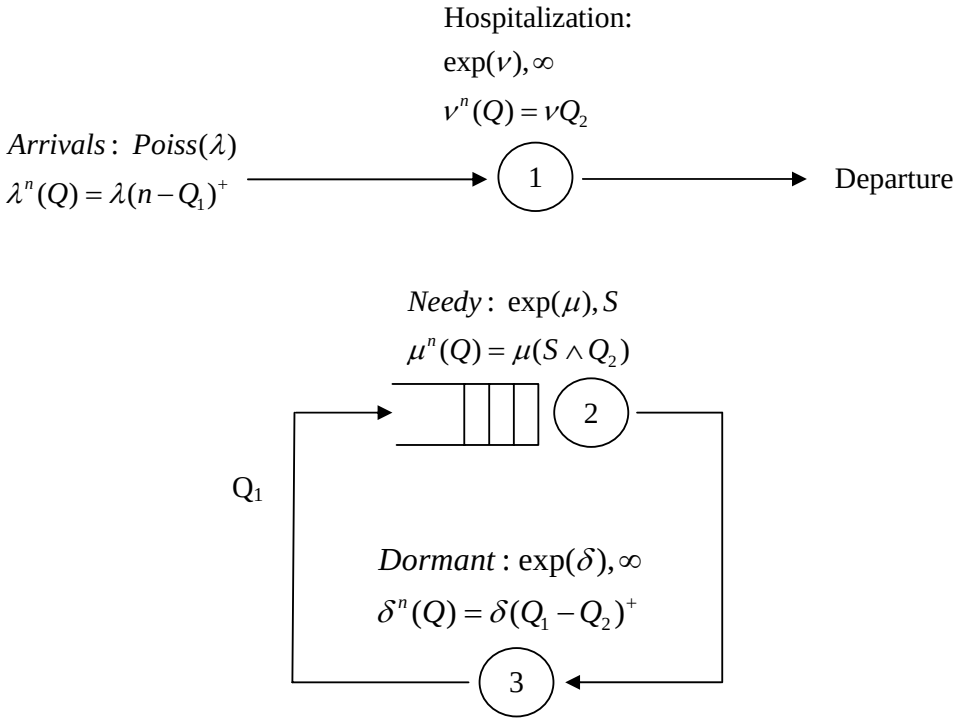


Figure 35: Alternative model - Proposal 4

12 Performance Measures

We now return to the first model, as stated in Chapter 11.2, and our further analysis is aimed exclusively at this.

12.1 Probability of Blocking

From the stationary probability, we will now deduce the probability P_l that there are l beds occupied in the system ($0 \leq l \leq n$). The beds could be occupied by patients in needy or dormant states or in the cleaning state. We will use the following relation

$$P_l := \sum_{\substack{i,j,k\geq 0 \\ i+j+k=l}} \pi(i,j,k) = \sum_{i=0}^{l}\sum_{j=0}^{l-i} \pi(i,j,l-i-j).$$

We distinguish two cases:

1. $l \leq s$:

$$\begin{aligned} P_l &= \pi_0 \sum_{i=0}^{l}\sum_{j=0}^{l-i} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j} \\ &= \pi_0 \frac{1}{l!}\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l \end{aligned}$$

2. $l > s$:

$$\begin{aligned} P_l &= \pi_0 \textstyle\sum_{i=0}^{l}\sum_{j=0}^{l-i} \frac{1}{\nu(i)}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j} \\ &= \pi_0 \left(\textstyle\sum_{i=0}^{s}\sum_{j=0}^{l-i} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j}\right. \\ &+ \left.\textstyle\sum_{i=s+1}^{l}\sum_{j=0}^{l-i} \frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j}\right) \\ &= \pi_0 \left(\frac{1}{l!}\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l\right. \\ &+ \left.\textstyle\sum_{i=s+1}^{l}\sum_{j=0}^{l-i} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right)\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j}\right) \end{aligned}$$

Thus,

$$\begin{aligned} P_l = \pi_0 \left(\frac{1}{l!}\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l \right. \\ \left. + I_{\{l>s\}} \sum_{i=s+1}^{l}\sum_{j=0}^{l-i} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right)\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(l-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{l-i-j}\right), \end{aligned} \tag{12.1}$$

where $I_{\{l>s\}}$ is the indicator function.

One can derive from that expression the quantity P_n, which is the probability of blocking of the medical unit. P_n will also be indicated as $P(blocked)$.

12.2 Probability of Waiting More Than t Units of Time and the Expected Waiting Time

One of the important parameters of the level of service, is the time spent in-queue. This is the time that a patient may have to wait to be treated. If a patient becomes needy when there are already i other needy patients in the unit, he will need to wait an in-queue random waiting time that follows an Erlang distribution with $(i - s + 1)^+$ stages, each with rate $s\mu$. The probability that this Erlang-distributed random variable is greater than t is $e^{-s\mu t}\sum_{j=0}^{i-s}(s\mu t)^j/(j!)$. Clearly, the patient only waits if $i \geq s$.

Let W denote the steady state, in-queue waiting time for a hypothetical patient, who just become needy, and denote $p_{n,s}(t)$ as the tail of the steady state distribution of W, given n beds and s nurses. Formally, $p_{n,s}(t) = P(W > t)$. As a consequence of dealing with a closed system, the total activation rate, i.e. the rate at which the collective stable patient population produces needy patients, is modulated by the state of the system, i.e., by the number of needy, dormant and cleaning beds. In addition, in order to calculate the tail of the steady state distribution of W, we need to use the Arrival Theorem for closed networks, quoted here from Chen and Yao [17].

> **The Arrival Theorem.** *In a closed Jackson network, the arrival at (or the departure from) any node observes time averages, with the job itself excluded. In particular, the probability that the network is in state*[2] $x - e_i$ *immediately before an arrival (or immediately after a departure) epoch at node i is equal to the ergodic distribution, of a closed network with one fewer job, in state $x - e_i$.*

Let $\pi^A(x - e_i)$, denote the probability that the system is in state $x - e_i$ at the arrival epoch of a customer to node i. Thus, immediately *after* the arrival of a customer to node i, the state is x. Then the arriving customer sees before him the state $x - e_i$, which corresponds to a network with one fewer job. Then by the arrival theorem, we conclude that $\pi^A(x - e_i) = \pi_{n-1}(x - e_i)$. In particular for the needy state (node 1), $\pi^A(x - e_1) = \pi_{n-1}(x - e_1) = \pi_{n-1}(i - 1, j, k)$.

The probability that a patient will get service immediately as he become needy is the sum of probabilities that the customer arriving at the needy state will see fewer than s needy patients; by the former notations it is equal to:

$$P(W = 0) = \sum_{l=0}^{n-1}\sum_{m=0}^{l}\sum_{i=0}^{min\{m,s-1\}} \pi^A(i, m - i, l - m). \tag{12.2}$$

[2] [17] refers to state x, rather than $x - e_i$ as we do; we believe [17] has a typo.

The distribution function of the waiting time is:

$$P(W \leq t) = P(W = 0) + \sum_{i=s}^{n-1} P(\text{there are } (i-s+1) \text{ patients who ended}$$

their service on time $\leq t$|Arrival at the needy state found i needy patients)$\cdot$

$$\cdot \pi^A(i, m-i, l-m) =$$

$$= P(W=0) + \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi^A(i, m-i, l-m) \int_0^t \frac{\mu s(\mu s x)^{i-s}}{(i-s)!} e^{-\mu s x} dx$$

$$= P(W=0) + \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi^A(i, m-i, l-m)(1 - \sum_{h=0}^{i-s} \frac{(\mu s t)^h}{h!} e^{-\mu s t}) \tag{12.3}$$

$$= \sum_{l=0}^{n-1} \sum_{m=0}^{l} \sum_{i=0}^{min\{m, s-1\}} \pi^A(i, m-i, l-m) + \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi^A(i, m-i, l-m)$$

$$- \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi^A(i, m-i, l-m) \sum_{h=0}^{i-s} \frac{(\mu s t)^h}{h!} e^{-\mu s t}$$

$$= 1 - \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m) \sum_{h=0}^{i-s} \frac{(\mu s t)^h}{h!} e^{-\mu s t}.$$

Therefore, the tail steady state distribution of W is

$$P(W > t) = \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m) \sum_{h=0}^{i-s} \frac{(\mu s t)^h}{h!} e^{-\mu s t} \tag{12.4}$$

and the expected waiting time E[W] can be derived via this tail formula, i.e.,

$$\begin{aligned} E[W] &= \int_0^\infty P(W > t) dt = \int_0^\infty \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m) \sum_{h=0}^{i-s} \frac{(\mu s t)^h}{h!} e^{-\mu s t} dt \\ &= \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m) \sum_{h=0}^{i-s} \int_0^\infty \frac{(\mu s t)^h}{h!} e^{-\mu s t} dt \\ &= \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m) \sum_{h=0}^{i-s} \frac{1}{\mu s} \\ &= \frac{1}{\mu s} \sum_{l=s}^{n-1} \sum_{m=s}^{l} \sum_{i=s}^{m} \pi_{n-1}(i, m-i, l-m)(i-s+1). \end{aligned} \tag{12.5}$$

This formula is exactly the same as the one found in Gross and Harris [38] pg. 193: In a closed Jackson network with M/M/c_j nodes, the mean waiting time at node j for a network containing n customers is $E(W_j(n)) = \frac{1}{\mu_j c_j} \sum_{i=c_j}^{n-1} (i - c_i + 1) p_j(i, n-1)$ where $p_j(i, n-1)$ is the marginal

probability of i in an $(n-1)$-customer system at node j. Therefore, for our system

$$\begin{aligned} E(W) &= \frac{1}{\mu s}\sum_{i=s}^{n-1}(i-s+1)p_1(i,n-1) \\ &= \frac{1}{\mu s}\sum_{i=s}^{n-1}(i-s+1)\sum_{l=i}^{n-1}\sum_{m=i}^{l}\pi_{n-1}(i,m-i,l-m) \\ &= \frac{1}{\mu s}\sum_{l=i}^{n-1}\sum_{m=i}^{l}\sum_{i=s}^{m}(i-s+1)\pi_{n-1}(i,m-i,l-m). \end{aligned} \tag{12.6}$$

12.3 Probability of Delay

The probability of delay in terms of previous definitions is $P(W>0)$. In order to find it we will again use the Arrival Theorem for closed networks, cited earlier on Page 71. Accordingly, we can derive performance measures of a medical unit with n beds and s nurses, by the steady-state distribution of the same system with $n-1$ beds and s nurses. The probability that a patient who becomes needy has to wait, is the probability that a patient will find more than s needy patients in a system with n beds, and this is exactly the steady-state probability of having more than s needy patients in a system with $n-1$ beds. By the notions of the arrival theorem, a patient entering node 1 (as he arrives) sees the system in state $x-e_i$ with probability $\pi_{n-1}(x-e_i)$ (no matter where he cames from). After his entrance the system state will be x. Therefore, if we want to know what is the probability that the patient will see the station full, we need to add up the probabilities that in the vector $x-e_1$ the first element x_1-1 will exceed s patients, i.e., we need to add together the arrival probabilities of all x such that $x_1-1\geq s$ and $|x|=\sum_i x_i \leq n$. Thus,

$$\begin{aligned} P(W>0) &= \sum_{x||x|\leq n;x_1-1\geq s}\pi^A(x-e_1) = \sum_{i,j,k|i+j+k\leq n-1,i\geq s}\pi^A(i,j,k) \\ &= \sum_{i\geq s}\pi_{n-1}(i,j,k) = \sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m). \end{aligned} \tag{12.7}$$

Thus, for the system with n beds and s nurses, the percentage of patients which are required to wait before being served, coincides with the probability that in a system with $n-1$ beds and s nurses, all the nurses are busy. Formally, $P(W>0)=P_{n-1}(N(\infty)\geq s)$.

12.4 Average Occupancy Level

The average occupancy level can be found by $OC(n,s)=\sum_{l=0}^{n}\sum_{m=0}^{l}\sum_{i=0}^{m}m\pi_n(i,m-i,l-m)$.

13 The QED Regime

Consider a sequence of s-server queues, indexed by n. Let the arrival-rates $\lambda_n \to \infty$, as $n \to \infty$, and fixed μ the service-rate. Define the offered load by $R_n = \frac{\lambda_n^{eff}}{\mu}$. The QED regime is achieved by choosing λ_n and s_n so that $\sqrt{s_n}(1-\rho_n) \to \beta$, as $n \to \infty$, for some finite β. Here $\rho_n = \frac{R_n}{s_n}$. When patients have infinite patience, ρ_n may be interpreted as the long-run servers' utilization and then one must have $0 < \beta < \infty$. Otherwise, ρ_n is the offered load per server and $-\infty < \beta < \infty$ is allowed. Equivalently, the staffing level is approximately given by

$$s_n \approx R_n + \beta\sqrt{R_n}, \qquad -\infty < \beta < \infty.$$

In our system $\lambda_n^{eff} = \frac{\lambda_n}{1-p}$, $R_n = \frac{\lambda_n}{(1-p)\mu}$, and $\rho_n = \frac{\lambda_n}{(1-p)s\mu}$.

Let λ, s and n tend to ∞ simultaneously so that:

$$s = \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}} + o(\sqrt{\lambda}), \qquad -\infty < \beta < \infty, \tag{i}$$

$$n - s = \eta_1\sqrt{\frac{\lambda}{(1-p)\mu}} + \frac{p\lambda}{(1-p)\delta} + \eta_2\sqrt{\frac{p\lambda}{(1-p)\delta}} + \frac{\lambda}{\gamma} + \eta_3\sqrt{\frac{\lambda}{\gamma}} + o(\sqrt{\lambda}), \tag{ii}$$

$$-\infty < \eta_1, \eta_2, \eta_3 < \infty.$$

First we reduce the number of parameters.

Theorem 7. *Let λ, s and n tend to ∞ simultaneously. Then the conditions*

$$s = \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}} + o(\sqrt{\lambda}), \qquad -\infty < \beta < \infty, \tag{i}$$

$$n - s = \eta_1\sqrt{\frac{\lambda}{(1-p)\mu}} + \frac{p\lambda}{(1-p)\delta} + \eta_2\sqrt{\frac{p\lambda}{(1-p)\delta}} + \frac{\lambda}{\gamma} + \eta_3\sqrt{\frac{\lambda}{\gamma}} + o(\sqrt{\lambda}), \tag{ii}$$

$$-\infty < \eta_1, \eta_2, \eta_3 < \infty$$

are equivalent to the conditions

$$\begin{aligned} &\text{(i) } s = \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}} + o(\sqrt{\lambda}), && -\infty < \beta < \infty \\ &\text{(ii) } n - s = \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} + o(\sqrt{\lambda}), && -\infty < \eta < \infty \end{aligned} \tag{13.1}$$

where $\eta = \eta_1\sqrt{\frac{\delta\gamma}{\mu(\gamma p+(1-p)\delta)}} + \eta_2\sqrt{\frac{\gamma p}{\gamma p+(1-p)\delta}} + \eta_3\sqrt{\frac{(1-p)\delta}{\gamma p+(1-p)\delta}}$.

Proof. Clearly, one can rewrite the second condition in the form

$$n - s = \left(\eta_1\sqrt{\frac{\delta\gamma}{\mu(\gamma p + (1-p)\delta)}} + \eta_2\sqrt{\frac{\gamma p}{\gamma p + (1-p)\delta}} + \eta_3\sqrt{\frac{(1-p)\delta}{\gamma p + (1-p)\delta}}\right)$$
$$\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + o(\sqrt{\lambda}), \quad -\infty < \eta_1, \eta_2, \eta_3 < \infty.$$

Setting $\eta = \eta_1\sqrt{\frac{\delta\gamma}{\mu(\gamma p+(1-p)\delta)}} + \eta_2\sqrt{\frac{\gamma p}{\gamma p+(1-p)\delta}} + \eta_3\sqrt{\frac{(1-p)\delta}{\gamma p+(1-p)\delta}}$ one obtains (13.1). The first condition is the same. This proves the statement. □

We can rewrite the QED condition (13.1) in the following form

$$\lim_{\lambda\to\infty} \frac{n - s - \frac{p\lambda}{(1-p)\delta} - \frac{\lambda}{\gamma}}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} = \eta, \qquad -\infty < \eta < \infty \qquad \text{(i)}$$

$$\lim_{\lambda\to\infty} \sqrt{s}\left(1 - \frac{\lambda}{(1-p)s\mu}\right) = \beta, \qquad -\infty < \beta < \infty \qquad \text{(ii)}$$

where the second term defines the situation on the servers (i.e. the effective space in the service station) and the first term defines the effective space remaining in the "non-queue" stations.

For convenience we denote

$$R_N = \frac{\lambda}{(1-p)\mu}, \quad R_D = \frac{p\lambda}{(1-p)\delta}, \quad R_C = \frac{\lambda}{\gamma},$$

and

$$\rho = \frac{\lambda}{(1-p)s\mu}.$$

For some technical reasons we must distinguish between two cases: $\beta = 0$ and $\beta \neq 0$. This separation results in two separate QED conditions:

$$QED = \begin{cases} \lim_{\lambda\to\infty} \frac{n-s-R_D-R_C}{\sqrt{R_D+R_C}} = \eta, & -\infty < \eta < \infty, & \text{(i)} \\ \lim_{\lambda\to\infty} \sqrt{s}\left(1 - \frac{R_N}{s}\right) = \beta, & -\infty < \beta < \infty,\ \beta \neq 0, & \text{(ii)} \end{cases} \tag{13.2}$$

and

$$QED_0 = \begin{cases} \lim_{\lambda\to\infty} \frac{n-s-R_D-R_C}{\sqrt{R_D+R_C}} = \eta, & -\infty < \eta < \infty, & \text{(i)} \\ \lim_{\lambda\to\infty} \sqrt{s}\left(1 - \frac{R_N}{s}\right) = \beta, & \beta = 0, & \text{(ii)} \end{cases}$$

where μ,p,δ and γ are fixed parameters.

14 Heavy Traffic Limits and Asymptotic Analysis in the QED Regime

In this chapter we develop heavy-traffic approximations of the system-measures introduced in Chapter 12. As a first stage we present four lemmas; their proofs are in Appendix C.2.

Lemma 1. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define ζ_1 as the expression*

$$\zeta_1 = \frac{e^{-R_N}}{s!}\left(R_N\right)^s \frac{1}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(R_D+R_C\right)^l e^{-(R_D+R_C)}.$$

Then

$$\lim_{\lambda\to\infty}\zeta_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta}.$$

Lemma 2. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define ζ_2 as the expression*

$$\zeta_2 = \frac{e^{-(R_N+R_D+R_C)}}{s!}\left(R_N\right)^s \frac{\rho^{n-s}}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{R_D}{\rho}+\frac{R_C}{\rho}\right)^l.$$

Then

$$\lim_{\lambda\to\infty}\zeta_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1).$$

Lemma 3. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED or QED_0 conditions. Define ξ as the expression*

$$\xi = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n-1}}\frac{1}{i!j!k!}\left(R_N\right)^i\left(R_D\right)^j\left(R_C\right)^k e^{-(R_N+R_D+R_C)}.$$

Then

$$\lim_{\lambda\to\infty}\xi = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t).$$

Lemma 4. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Define ζ as the expression*

$$\zeta = e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}{R_D}^j{R_C}^k\sum_{i=0}^{n-s-j-k-1}\rho^i.$$

Then

$$\lim_{\lambda\to\infty}\zeta = \sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right).$$

14.1 Approximation of the Probability of Delay

The first approximation will be for the measure: the probability of waiting or the probability of delay. It was defined in Section 12.3, by Formula (12.7).

Theorem 8. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Then*

$$\lim_{\lambda\to\infty} P(W>0) = \left(1+\frac{\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{B}\right)d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}\right)^{-1}$$

where $B=\frac{R_N}{R_C+R_D}=\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, $\eta_1=\eta-\beta\sqrt{B^{-1}}$.

Proof.

$$P_n(W>0)=P_{n-1}(Q1(\infty)\geq s)=\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m)$$
$$=\pi_0^{n-1}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\frac{1}{s!s^{i-s}(m-i)!(l-m)!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i}\left(\frac{\lambda}{\gamma}\right)^{l-m},$$

where

$$\pi_0^{n-1}=\left(\sum_{l=0}^{n-1}\frac{1}{l!}\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^l\right.$$
$$\left.+\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\left(\frac{1}{s!s^{i-s}}-\frac{1}{i!}\right)\frac{1}{(m-i)!(l-m)!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i}\left(\frac{\lambda}{\gamma}\right)^{l-m}\right)^{-1}.$$

Thus,

$$P_n(W>0)=\left(1+\frac{A}{B}\right)^{-1},$$

where

$$A=\sum_{l=0}^{n-1}\frac{1}{l!}\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^l$$
$$-\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\frac{1}{i!(m-i)!(l-m)!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i}\left(\frac{\lambda}{\gamma}\right)^{l-m} \tag{14.1}$$
$$=\sum_{\substack{i,j,k|i\leq s,\\ i+j+k\leq n-1}}\frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\left(\frac{p\lambda}{(1-p)\delta}\right)^j\left(\frac{\lambda}{\gamma}\right)^k,$$

$$
\begin{aligned}
B &= \sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\frac{1}{s!s^{i-s}}\frac{1}{(m-i)!(l-m)!}\left(\frac{\lambda}{(1-p)\mu}\right)^{i}\left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i}\left(\frac{\lambda}{\gamma}\right)^{l-m}\\
&= \sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\sum_{i=s}^{n-j-k-1}\frac{1}{s!s^{i-s}}\frac{1}{j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^{i}\left(\frac{p\lambda}{(1-p)\delta}\right)^{j}\left(\frac{\lambda}{\gamma}\right)^{k}\\
&= \sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\sum_{i=0}^{n-s-j-k-1}\frac{1}{s!s^{i}}\frac{1}{j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^{i+s}\left(\frac{p\lambda}{(1-p)\delta}\right)^{j}\left(\frac{\lambda}{\gamma}\right)^{k}\\
&= \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(\frac{p\lambda}{(1-p)\delta}\right)^{j}\left(\frac{\lambda}{\gamma}\right)^{k}\sum_{i=0}^{n-s-j-k-1}\left(\frac{\lambda}{(1-p)s\mu}\right)^{i}.
\end{aligned}
\tag{14.2}
$$

Define $\rho = \frac{\lambda}{(1-p)s\mu}$, then under the QED (part (ii)) assumption that $\sqrt{s}(1-\rho)\to\beta,\ -\infty<\beta<\infty, \beta\neq 0$ (of Theorem 8) as $\lambda\to\infty$, we can rewrite the right-hand side in the following way:

$$
\begin{aligned}
B &= \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(\frac{p\lambda}{(1-p)\delta}\right)^{j}\left(\frac{\lambda}{\gamma}\right)^{k}\frac{1-\rho^{n-s-j-k}}{1-\rho}\\
&= \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\frac{1}{1-\rho}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(\frac{p\lambda}{(1-p)\delta}\right)^{j}\left(\frac{\lambda}{\gamma}\right)^{k}\\
&- \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\frac{\rho^{n-s}}{1-\rho}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(\frac{p\lambda}{(1-p)\delta\rho}\right)^{j}\left(\frac{\lambda}{\gamma\rho}\right)^{k}.
\end{aligned}
$$

Applying the multinomial theorem yields:

$$
\begin{aligned}
B &= \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\frac{1}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^{l}\\
&- \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\frac{\rho^{n-s}}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)^{l}\\
&= B_1 - B_2.
\end{aligned}
$$

Multiplying A, B_1 and B_2 by $e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$ we have

$$
P(W>0) = \left(1+\frac{\xi}{\zeta_1-\zeta_2}\right)^{-1},
$$

where ξ, ζ_1 and ζ_2 where defined in lemmas 1-3, and we repeat them for convenience,

$$
\xi = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n-1}}\frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^{i}\left(\frac{p\lambda}{(1-p)\delta}\right)^{j}\left(\frac{\lambda}{\gamma}\right)^{k}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}, \tag{14.3}
$$

$$
\zeta_1 = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\frac{1}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^{l}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}, \tag{14.4}
$$

$$
\zeta_2 = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^{s}\frac{\rho^{n-s}}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)^{l}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}. \tag{14.5}
$$

By Lemmas 1,2, and 3 if $\beta \neq 0$:

$$\lim_{\lambda\to\infty} \zeta_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta}, \tag{14.6}$$

$$\lim_{\lambda\to\infty} \zeta_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1), \tag{14.7}$$

and

$$\lim_{\lambda\to\infty} \xi = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) d\Phi(t), \tag{14.8}$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}$. Thus,

$$\lim_{\lambda\to\infty} P(W>0) = \left(1 + \frac{\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}\right)^{-1}.$$

This proves Theorem 8. □

Theorem 9. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Then*

$$\lim_{\lambda\to\infty} P(W>0) = \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) d\Phi(t)}{\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right)}\right)^{-1}$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}$.

Proof. As before,

$$P_n(W>0) = \left(1 + \frac{A}{B}\right)^{-1}$$

where,

$$A = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n-1}} \frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\gamma}\right)^k$$

$$B = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1} \frac{1}{j!k!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\gamma}\right)^k \sum_{i=0}^{n-s-j-k-1}\left(\frac{\lambda}{(1-p)s\mu}\right)^i.$$

We can multiply each phrase in $e^{-(R_N+R_D+R_C)}$ where $R_N = \frac{\lambda}{(1-p)\mu}$, $R_D = \frac{p\lambda}{(1-p)\delta}$, $R_C = \frac{\lambda}{\gamma}$, then

$$P_n(W>0) = \left(1 + \frac{\xi}{\zeta}\right)^{-1}$$

where, $\xi = A \cdot e^{-(R_N+R_D+R_C)}$, and $\zeta = B \cdot e^{-(R_N+R_D+R_C)}$. By Lemma 3, when $\beta = 0$:

$$\lim_{\lambda\to\infty} \xi = \int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) d\Phi(t), \tag{14.9}$$

and, due to Lemma 4:

$$\lim_{\lambda\to\infty} \zeta = \sqrt{\frac{(1-p)\mu}{\gamma} + \frac{p\mu}{\delta}} \frac{1}{\sqrt{2\pi}} \left(\eta\Phi(\eta) + \phi(\eta)\right). \tag{14.10}$$

Assigning Equations (14.9) and (14.10), we proved Theorem 9.

□

Checking: Is $\lim_{\beta\to 0} P_{\{\beta\neq 0\}}(W > 0) = P_{\{\beta=0\}}(W > 0)$?

We need to check that:

$$\lim_{\beta\to 0} \frac{\phi(\beta)\Phi(\eta) - \phi(\sqrt{\eta^2+\beta^2})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}{\beta}$$
$$= \sqrt{\frac{(1-p)\mu}{\gamma} + \frac{p\mu}{\delta}} \frac{1}{\sqrt{2\pi}} \left(\eta\Phi(\eta) + \phi(\eta)\right).$$

Define: $\eta_1 = \eta - \beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} = \eta - \beta\sqrt{C}$. By L'Hôpital's rule:

$$\lim_{\beta\to 0} \frac{\phi(\beta)\Phi(\eta) - \phi(\sqrt{\eta^2+\beta^2})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}{\beta}$$
$$= \lim_{\beta\to 0} \frac{d}{d\beta}\left(\phi(\beta)\Phi(\eta) - \phi(\sqrt{\eta^2+\beta^2})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\right)$$
$$= \lim_{\beta\to 0} \Phi(\eta)\frac{d\phi(\beta)}{d\beta} - \frac{d\phi(\sqrt{\eta^2+\beta^2})}{d\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1) - \phi(\sqrt{\eta^2+\beta^2})\frac{d(e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1))}{d\beta}$$
$$= \lim_{\beta\to 0} \Phi(\eta)\frac{d\phi(\beta)}{d\beta} - \frac{d\phi(\sqrt{\eta^2+\beta^2})}{d\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1) - \phi(\sqrt{\eta^2+\beta^2})\left(\frac{de^{\frac{1}{2}\eta_1^2}}{d\beta}\Phi(\eta_1) + \frac{d\Phi(\eta_1)}{d\beta}e^{\frac{1}{2}\eta_1^2}\right)$$
$$= \lim_{\beta\to 0} \Phi(\eta)\frac{-\beta}{\sqrt{2\pi}}e^{-\frac{\beta^2}{2}} - \frac{-\beta}{\sqrt{2\pi}}e^{-\frac{\eta^2+\beta^2}{2}}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)$$
$$- \phi(\sqrt{\eta^2+\beta^2})\left(\frac{d\left(\frac{1}{2}\eta_1^2\right)}{d\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1) + \frac{d\eta_1}{d\beta}\phi(\eta_1)e^{\frac{1}{2}\eta_1^2}\right)$$
$$= \lim_{\beta\to 0} \Phi(\eta)\frac{-\beta}{\sqrt{2\pi}}e^{-\frac{\beta^2}{2}} - \frac{-\beta}{\sqrt{2\pi}}e^{-\frac{\eta^2+\beta^2}{2}}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)$$
$$- \phi(\sqrt{\eta^2+\beta^2})\left((-\eta_1\sqrt{C})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1) + (-\sqrt{C})\phi(\eta_1)e^{\frac{1}{2}\eta_1^2}\right)$$
$$= \phi(\eta)\sqrt{C}e^{\frac{1}{2}\eta^2}\left(\eta\Phi(\eta) + \phi(\eta)\right) = \frac{\sqrt{C}}{\sqrt{2\pi}}\left(\eta\Phi(\eta) + \phi(\eta)\right).$$

14.2 Approximation of the Expected Waiting Time

The second approximation will be for the measure: the expected waiting time. It was defined in Section 12.2, by Formula (12.5). We state the Theorems here, the proofs are in Appendix C.3.

The first theorem gives the approximation for the case where $\beta \neq 0$.

Theorem 10. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Then*

$$\lim_{\lambda\to\infty} \sqrt{s}E[W] = \frac{1}{\mu} \frac{\frac{\phi(\beta)\Phi(\eta)}{\beta}\frac{1}{\beta} + \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\left(\frac{\beta}{B} - \frac{\eta}{\sqrt{B}} - \frac{1}{\beta}\right)}{\int_{-\infty}^{\beta}\Phi\left(\eta + (\beta - t)\sqrt{B}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}$$

where $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, $\eta_1 = \eta - \beta\sqrt{B^{-1}}$.

The second theorem gives the approximation for the case where $\beta = 0$.

Theorem 11. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Then*

$$\lim_{\lambda\to\infty} \sqrt{s}E[W] = \frac{1}{2\mu} \frac{B^{-1}\left((\eta^2+1)\Phi(\eta) + \eta\phi(\eta)\right)}{\sqrt{2\pi}\int_{-\infty}^{0}\Phi\left(\eta - t\sqrt{B}\right)d\Phi(t) + \sqrt{B^{-1}}\left(\eta\Phi(\eta) + \phi(\eta)\right)}$$

where $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, $\eta_1 = \eta - \beta\sqrt{B^{-1}}$.

14.3 Approximation of the Blocking Probability

The third approximation will be for the probability of blocking. This measure was defined in Section 12.1, by Formula (12.1). We only state here the approximation theorems, as conjectures supported by our previous experience. We intend to prove these measures in the near future.

Theorem 12. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define* $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, *then*

$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2) + \phi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi\left(\eta + (\beta - t)\sqrt{B}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)} \tag{14.11}$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$, $\nu = \frac{1}{\sqrt{1+B^{-1}}}$, $\nu_1 = \frac{\eta\sqrt{B^{-1}}+\beta}{\sqrt{1+B^{-1}}}$, $\nu_2 = \frac{\beta\sqrt{B^{-1}}-\eta}{\sqrt{1+B^{-1}}}$.

Theorem 13. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Define and* $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, *then*

$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{0}\Phi\left(\eta - t\sqrt{B}\right)d\Phi(t) + \frac{1}{\sqrt{B}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta) + \phi(\eta)\right)} \tag{14.12}$$

where $\nu = \frac{1}{\sqrt{1+B^{-1}}}$, $\nu_1 = \frac{\eta}{\sqrt{1+B}}$, $\nu_2 = \frac{\eta}{\sqrt{1+B^{-1}}}$.

15 Comparison of Approximations and Exact Calculations

In this section, using some examples, we illustrate the quality of our approximations, when compared against the exact calculations. We answer the question of what are the parameter settings (e.g. what β) for which we can use the approximations. The exact calculations and the approximations were calculated using MATLAB. We could compare only systems where n is less than 160. Larger systems are beyond MATLAB capabilities and the exact formulas cannot be calculated. Nevertheless, since our asymptotes become more accurate as n and s grows to infinity together, this limitation only strengthens our analysis.

In order to simplify this experimental phase, we will assume that $\gamma = \infty$, which means that we neglect the Cleaning phase and actually consider the semi-open Erlang-R system (see Figure 28).

We considered various combinations of the data that cover small, moderate and large systems, and a range of values of the offered-load ratio. The parameters of each case is specified in Table 2. For each set of parameters we calculated s and n when $\beta \in [-3, 3]$ and $\eta \in [-3, 3]$. The following graphs compare visually the exact and the approximate calculations.

Figure	n	s	λ	δ	μ	p	B
36	77-162	71-131	10	5	1	0.9	0.55
37	60-150	25-66	15	0.5	1	0.667	0.75
38	51-133	29-72	30	0.5	1	0.4	1.25
39	6-35	1-15	10	1	3	0.5	0.66
40	1-30	2-15	5	0.25	1	0.5	1

Table 2: Parameters for a large system

The first illustration, shown in Figure 36, is of a large system where the number of "beds" is up to 162, and the number of servers is up to 131. We observe an excellent matching between the exact calculation and the approximation for the $P(W > 0)$ measure. We also observe a very good match for the P(block) measure when $\beta > -0.5$. As β decreases, the approximations becomes less accurate. This is expected since, as β decreases, we exit the QED regime. For example, when $\beta = -2$, $\sqrt{s} * P(block) \approx 2$. In this case, if $s \approx 100$ then $P(block) \approx 20\%$. But the QED regime prevails when P(block) is less then 10%, and $P(W > 0)$ is in $[25\%, 75\%]$.

The next illustrations, shown in Figure 37 and 38, are of medium size systems, where the number of servers is between 20 to 70. Here too, we observe a very good matching for the probability of waiting. In the P(block) graph, the pattern is more complex. We observe a very good match when

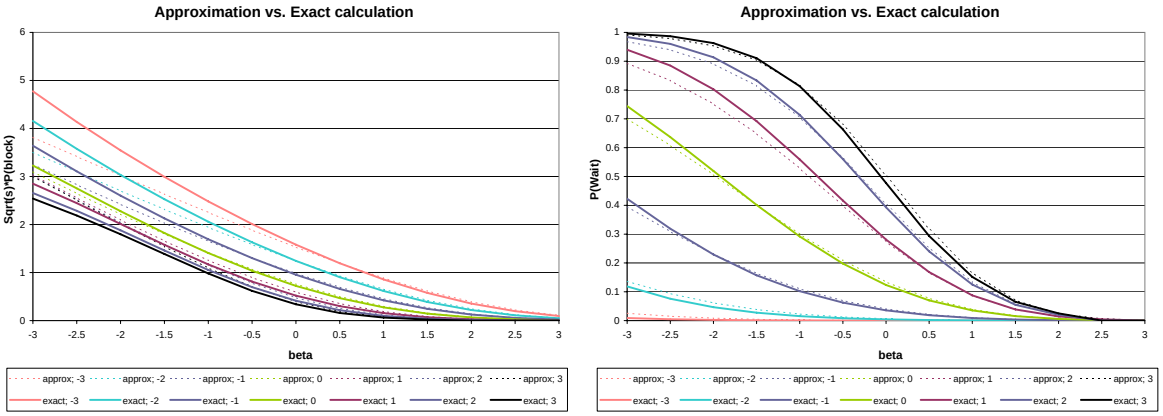


Figure 36: Comparison of approximation and exact calculation - Large system

$\beta > -0.5$ and $\eta > -1$. As η and β get smaller, the approximations become less accurate. We also see that as η get smaller, the value of β from which the approximations are accurate is larger.

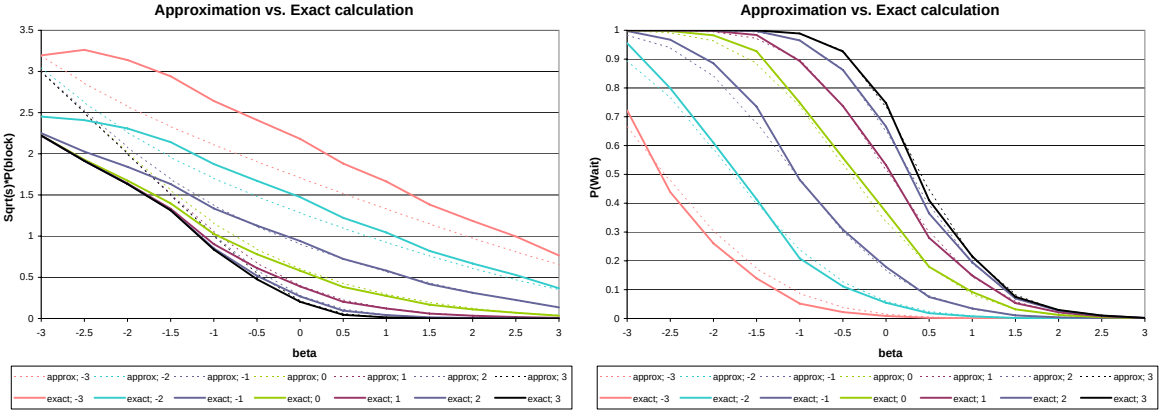


Figure 37: Comparison of approximation and exact calculation - Medium system

The next two illustrations, shown in Figure 39 and 40, are of small size systems where the number of servers is up to 15, and the number of beds is less than 35. We observe the same phenomena here as in medium systems, but we also see that there are additional inaccuracies in the $P(W > 0)$ approximation for $\eta < 0$.

From this experiment, we conclude that, while in the QED regime the approximations are remarkably accurate, when exiting this operating regime, the $P(W > 0)$ approximation remains stable, but the P(block) approximation does not.

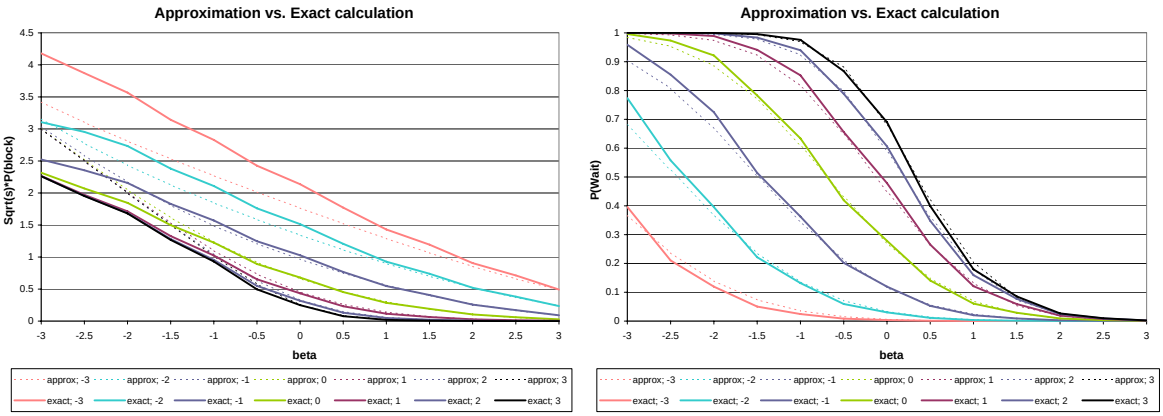


Figure 38: Comparison of approximation and exact calculation - Medium system

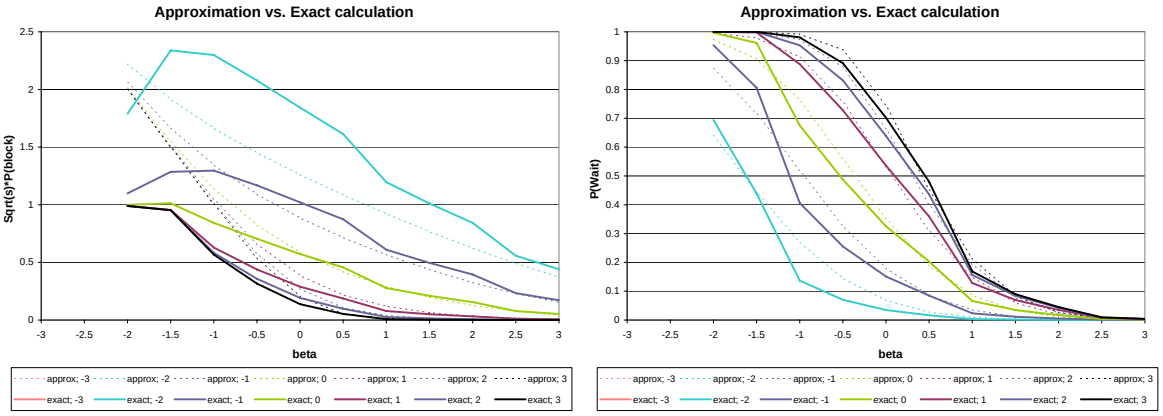


Figure 39: Comparison of approximation and exact calculation - Small system

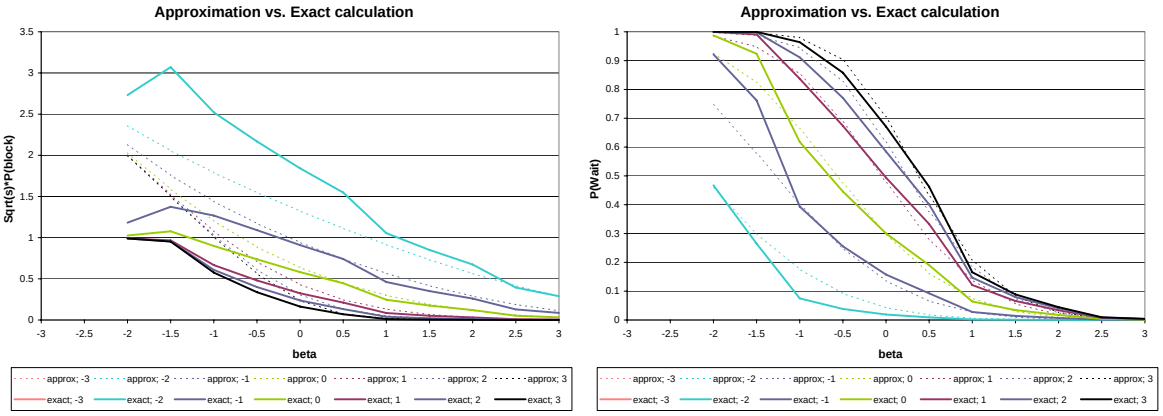


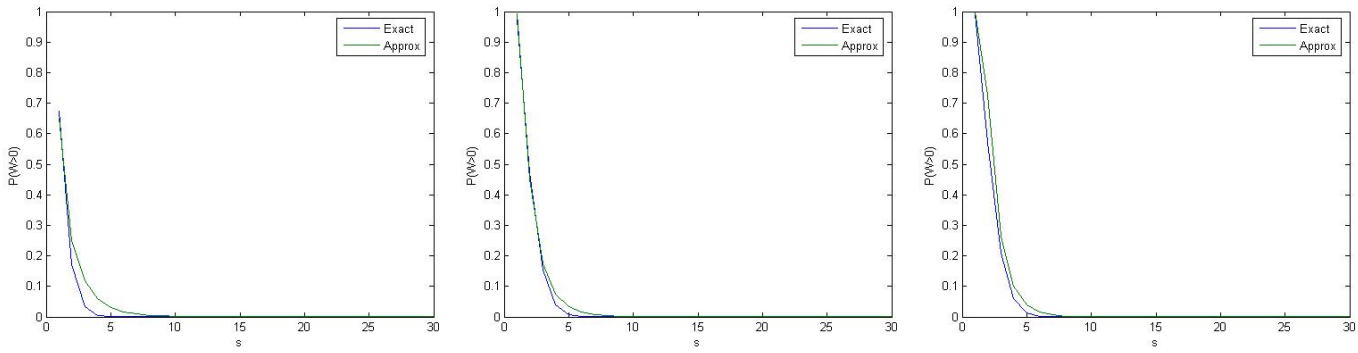
Figure 40: Comparison of approximation and exact calculation - Small system

The next three experiment-parameters are based on data taken from surveys carried out in Israeli Hospitals by Marmor [57]. The data we have are partial, not altogether consistent with the data we need. We know that in each of these MUs, there are 30 beds, and the average LOS is three days; patients' average arrival rate is five patients per day. Since we do not have accurate data on service times, we show three options using various assumptions. In all of them, we assume that average cleaning times are one hour. The parameters are specified in Table 3 and their related Figure is 41. In all the Figures (41 a-c), we observe a good matching between the exact calculation and the

Figure	n	λ	δ	μ	γ	p
41-a	30	5/24	4/23	4	1	12/13
41-b	30	5/24	4/21	4	1	36/37
41-c	30	5/24	2/19	2	1	30/31

Table 3: Parameters based on data from Israeli hospital

approximation in the range of interest (i.e. when $P(W > 0) \in [0.3 - 0.7]$).



(a) (b) (c)

Figure 41: Comparison of approximation and exact calculation - Israeli Hospital

The last experiment uses parameters taken from the work of Jennings and de Véricourt, so that their model can be compared to ours. They used the following ratio: $r = \frac{\lambda_J}{\lambda_J+\mu} = \frac{\delta}{\delta+\mu} = 0.25$. The illustration is shown in Figure 42; the parameters are specified in Table 4. As in previous cases, we

Figure	n	λ	δ	μ	γ	p
42	40	10	1	3	1	0.5

Table 4: Parameters based on Jennings and de Véricourt's article

observe a very good matching between the exact calculation and the approximation.

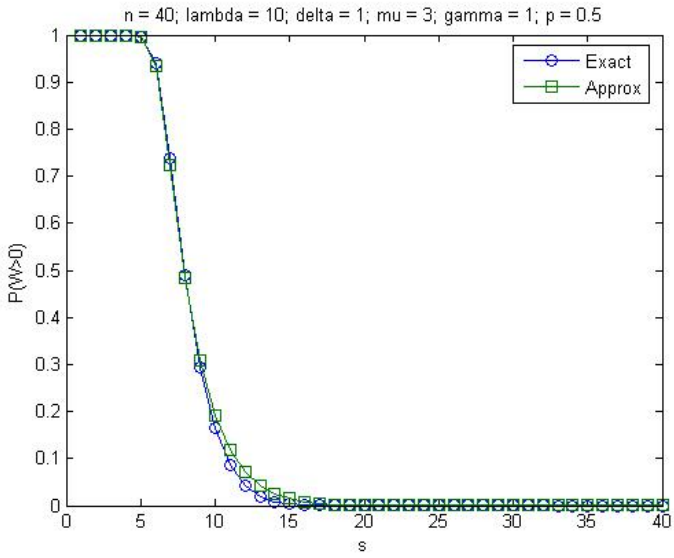


Figure 42: Comparison of approximation and exact calculation - $r = 0.25$

16 Comparison with Other Models

In this chapter we will present some special cases of our model. We will show that the probability distribution of these models can be represented by our model probability functions, or some other connection between the models.

16.1 The M/M/s/infinity/n System

When $\lambda \to \infty$, and $\delta = \gamma$ our model will be equivalent to Jennings and de Véricourt's model [46]. This is an M/M/S/∞/n system, with exactly n customer in the system, i.e. $i+j+k=n$. Jennings and de Véricourt used the definition of $r = \frac{\lambda_J}{\lambda_J+\mu}$ which is equivalent to: $r = \frac{p\delta+(1-p)\gamma}{p\delta+(1-p)\gamma+\mu} = \frac{\delta}{\delta+\mu}$ in our model.

As seen before, for each (i,j,k) such that $0 \le i+j+k \le n$,

$$\pi(i,j,k) = \pi_0 \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\delta}\right)^k$$

since $\lambda = \infty$ (goes to infinity faster than n and s) the probability of patients being in node 4 is 0, thus, $i+j+k=n$ and $\pi(i,j,k) = \pi(i,j,n-i-j)$. We define the marginal distribution $\pi(i,n-i)$ as $j+k=n-i$:

$$\begin{aligned}
\pi(i,n-i) &= \sum_{j,k|j+k=n-i} \pi(i,j,k) \\
&= \sum_{j,k|j+k=n-i} \pi_0 \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!k!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\delta}\right)^k \\
&= \pi_0 (\lambda)^n \frac{1}{\nu(i)} \left(\frac{1}{(1-p)\mu}\right)^i \frac{1}{(n-i)!} \left(\frac{1}{(1-p)\delta}\right)^{n-i} \\
&= \pi_0 \left(\frac{\lambda}{1-p}\right)^n \frac{1}{\nu(i)} \left(\frac{1}{\mu}\right)^i \frac{1}{(n-i)!} \left(\frac{1}{\delta}\right)^{n-i} \\
&= \pi_0 \left(\frac{\lambda}{(1-p)\delta}\right)^n \frac{1}{\nu(i)} \frac{1}{(n-i)!} \left(\frac{\delta}{\mu}\right)^i.
\end{aligned}$$

Here $\nu(i)$ is defined as

$$\nu(i) := \begin{cases} i! & , \quad i \le s, \\ s!s^{i-s} & , \quad i \ge s. \end{cases}$$

and π_0 is given by

$$\begin{aligned}
\pi_0^{-1} &= \sum_{i+j+k=n} \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k \\
&= \lambda^n \sum_{i=0}^{n} \sum_{j,k|j+k=n-i} \frac{1}{\nu(i)} \left(\frac{1}{(1-p)\mu}\right)^i \frac{1}{j!k!} \left(\frac{p}{(1-p)\delta}\right)^j \left(\frac{1}{\delta}\right)^k \\
&= \lambda^n \sum_{i=0}^{n} \frac{1}{\nu(i)} \left(\frac{1}{(1-p)\mu}\right)^i \frac{1}{(n-i)!} \left(\frac{1}{(1-p)\delta}\right)^{n-i} \\
&= \left(\frac{\lambda}{1-p}\right)^n \left(\sum_{i=0}^{s} \frac{1}{i!} \left(\frac{1}{\mu}\right)^i \frac{1}{(n-i)!} \left(\frac{1}{\delta}\right)^{n-i} + \sum_{i=s+1}^{n} \frac{1}{s!s^{i-s}} \left(\frac{1}{\mu}\right)^i \frac{1}{(n-i)!} \left(\frac{1}{\delta}\right)^{n-i} \right) \\
&= \left(\frac{\lambda}{1-p}\right)^n \left(\left(\frac{1}{\mu}+\frac{1}{\delta}\right)^n + \sum_{i=s+1}^{n} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \frac{n!}{(n-i)!} \left(\frac{1}{\mu}\right)^i \left(\frac{1}{\delta}\right)^{n-i} \right).
\end{aligned}$$

Or in an equivalent form,

$$\pi(i,n-i) = \begin{cases} \tilde{\pi}_0 \dfrac{1}{i!} \left(\dfrac{1}{\mu}\right)^i \dfrac{1}{(n-i)!} \left(\dfrac{1}{\delta}\right)^{n-i}, & i \leq s, \\ \tilde{\pi}_0 \dfrac{1}{s!s^{i-s}} \left(\dfrac{1}{\mu}\right)^i \dfrac{1}{(n-i)!} \left(\dfrac{1}{\delta}\right)^{n-i}, & i > s, \\ 0, & otherwise, \end{cases}$$

where $\tilde{\pi}_0$ is given by

$$\tilde{\pi}_0^{-1} = \left(\frac{1}{\mu}+\frac{1}{\delta}\right)^n + \sum_{i=s+1}^{n} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \frac{n!}{(n-i)!} \left(\frac{1}{\mu}\right)^i \left(\frac{1}{\delta}\right)^{n-i}.$$

Or in an equivalent form,

$$\pi(i,n-i) = \begin{cases} \bar{\pi}_0 \binom{n}{i} \left(\dfrac{\delta}{\mu}\right)^i, & i \leq s, \\ \bar{\pi}_0 \binom{n}{i} \dfrac{i!}{s!s^{i-s}} \left(\dfrac{\delta}{\mu}\right)^i, & i > s, \\ 0, & otherwise, \end{cases}$$

where $\bar{\pi}_0$ is given by

$$\begin{aligned}
\bar{\pi}_0^{-1} &= n!\delta^n \left(\left(\frac{1}{\mu}+\frac{1}{\delta}\right)^n + \sum_{i=s+1}^{n} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \frac{n!}{(n-i)!} \left(\frac{1}{\mu}\right)^i \left(\frac{1}{\delta}\right)^{n-i} \right) \\
&= n! \left(\left(\frac{\mu+\delta}{\mu}\right)^n + \sum_{i=s+1}^{n} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \frac{n!}{(n-i)!} \left(\frac{\delta}{\mu}\right)^i \right).
\end{aligned}$$

This last version of steady-state probabilities were investigated by Jennings and de Véricourt [46]. In [46] there are no exogenous arrivals (closed system i.e. no λ in our notation), which corresponds to taking λ to infinity at a faster rate than n and s.

16.2 Call Center with IVR (Interactive Voice Response)

In certain settings, our model will be equivalent to Khudyakov's model [49]. As seen before, for each (i,j,k) such that $0 \leq i+j+k \leq n$:

$$\pi(i,j,k) = \pi_0 \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k.$$

We would like to define the marginal distribution $\pi(i,l)$ as $j+k=l$. For each (i,l) such as $0 \leq i+l \leq n$:

$$\begin{aligned}
\pi(i,l) &= \sum_{j,k|j+k=l} \pi(i,j,k) \\
&= \sum_{j,k|j+k=l} \pi_0 \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!k!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\gamma}\right)^k \\
&= \pi_0 \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{l!} \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l.
\end{aligned}$$

Here $\nu(i)$ is defined as

$$\nu(i) := \begin{cases} i! & , \quad i \leq s, \\ s!s^{i-s} & , \quad i \geq s, \end{cases}$$

where π_0 is given by

$$\begin{aligned}
\pi_0^{-1} &= \sum_{0 \leq i+j+k \leq n} \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k \\
&= \sum_{i=0}^{n} \sum_{0 \leq j+k \leq n-i} \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k \\
&= \sum_{i=0}^{n} \sum_{l=0}^{n-i} \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{l!} \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l \\
&= \sum_{i \leq s, i+l \leq n} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{l!} \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l \\
&\qquad + \sum_{i > s, i+l \leq n} \frac{1}{s!s^{i-s}} \left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{l!} \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l.
\end{aligned}$$

These are exactly the steady-state probabilities investigated by Khudyakov [49], in the case: $\frac{1}{\theta_K} = \left(\frac{p}{(1-p)\delta} + \frac{1}{\gamma}\right)$ and $\frac{p_K}{\mu_K} = \frac{1}{(1-p)\mu}$. For example: if in our model $p=0$, our model is comparable to Khudyakov's model with $p_K = 1$.

17 Generalizations

In previous chapters we decided, arbitrarily, that patients will start and end their stay at the MU, in service from nurses. One can use other assumptions, which will slightly change the shape of the network. In this chapter, we would like to generalize our findings, and to formulate heavy-traffic approximations that will cover any specific design of flow in the network. In fact, this generalization will also unite our model with the IVR model [49]. The generalization was developed from our experience in developing the approximations in Chapter 14. We have not yet proved them, and we might do so if we find it useful for the later stages of our research. This generalization covers any semi-open network, with one service station with s servers, and any finite number of delay procedures.

Consider any closed Jackson network that has one M/M/S node (denote as node a), any K finite number of nodes of the type M/M/∞ (denote as node j , $j \in J = \{1, 2, ..., K\}$) and one node of the type M/M/1. Denote the solution of the balance equations of the steady-state flows of the network as ρ_a and ρ_j, $\forall j \in J$, and define $A = \sum_{j \in J} \rho_j$, $B = \frac{\rho_a}{A}$. (Explanation: Let λ be the service rate of the M/M/1 node, and P be the transition probability matrix of the network. Define λ_j as the rate of flow into node j and μ_j as the rate of flow out of node j, then $\rho_j = \frac{\lambda_j}{\mu_j}$. The rate λ_j can be found by solving the traffic equations, and should be expressed in terms of λ, and P). The steady-state probabilities of such networks are:

$$\pi(i_0, i_1, ..., i_K) = \begin{cases} \pi_0 \frac{1}{\nu(i_0)} (\rho_a)^{i_0} \prod_{j=1}^{K} \frac{1}{i_j!} (\rho_j)^{i_j} , & 0 \leq i + \sum_{j=1}^{K} i_j \leq n, \\ 0 & otherwise. \end{cases}$$

Here $\nu(i)$ is defined as

$$\nu(i_0) := \begin{cases} i_0! & , \quad i_0 \leq s, \\ s! s^{i_0 - s} & , \quad i_0 \geq s, \end{cases}$$

and π_0 is the normalization factor, given by

$$\begin{aligned} \pi_0^{-1} &= \sum_{0 \leq i_0 + i_1 + ... + i_K \leq n} \left(\frac{1}{\nu(i_0)} (\rho_a)^{i_0} \prod_{j \in J} \frac{1}{i_j!} (\rho_j)^{i_j} \right) \\ &= \sum_{i=0}^{n} \sum_{l=0}^{n-i} \left(\frac{1}{\nu(i)} (\rho_a)^{i} \frac{1}{l!} (A)^{l} \right). \end{aligned}$$

We claim that for some service-level objectives the steady-state probabilities have the same structure as those investigated by Khudyakov [49] and in this work. For example, one can develop the marginal distribution, the probability of delay and blocking, and $E[W]$.

17.1 The Marginal Distribution

We will calculate the marginal distribution $\pi(i,l)$ as $l=\sum_{j=1}^{K} i_j$. For each (i,l) such as $0 \leq i+l \leq n$:

$$\begin{aligned}\pi(i,l) &= \sum_{i_1,\ldots,i_K | i_1+\ldots+i_K=l} \pi(i,i_1,..,i_K) \\ &= \sum_{i_1,\ldots,i_K | i_1+\ldots+i_K=l} \left(\pi_0 \frac{1}{\nu(i)} (\rho_a)^i \prod_{j\in J} \frac{1}{i_j!} (\rho_j)^{i_j} \right) \\ &= \pi_0 \frac{1}{\nu(i)} (\rho_a)^i \frac{1}{l!} (A)^l .\end{aligned}$$

17.2 The Probability of Delay

The second example we show is for the calculation of the probability of delay:

$$\begin{aligned}P_n(W>0) = P_{n-1}(Q1(\infty) \geq s) &= \sum_{s\leq i+i_1+\ldots+i_K \leq n-1 | i \geq s} \pi_{n-1}(i,i_1,...,i_K) \\ = \pi_0 \sum_{s\leq i+i_1+\ldots+i_K \leq n-1 | i \geq s} & \frac{1}{s!s^{i-s}} (\rho_a)^i \prod_{j\in J} \frac{1}{i_j!} (\rho_j)^{i_j}\end{aligned} \tag{17.1}$$

where,

$$\begin{aligned}\pi_0^{-1} &= \sum_{0\leq i_0+i_1+\ldots+i_K \leq n-1} \left(\frac{1}{\nu(i_0)} (\rho_a)^{i_0} \prod_{j\in J} \frac{1}{i_j!} (\rho_j)^{i_j} \right) \\ &= \sum_{i=0}^{n-1} \sum_{l=0}^{n-i-1} \left(\frac{1}{\nu(i)} (\rho_a)^i \frac{1}{l!} (A)^l \right).\end{aligned}$$

These probabilities have the same structure as that which was investigated in Chapter 14. Therefore, we can conclude with the following general theorems:

Theorem 14. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty} \frac{n-s-A}{\sqrt{A}} = \eta, \qquad -\infty < \eta < \infty, \tag{i}$$

$$\lim_{\lambda\to\infty} \sqrt{s}\left(1-\frac{\rho_a}{s}\right) = \beta, \qquad -\infty < \beta < \infty,\ \beta \neq 0 \tag{ii}$$

where all other parameters are fixed. Then

$$\lim_{\lambda\to\infty} P(W>0) = \left(1 + \frac{\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta-t)\sqrt{B}\right) d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2} \Phi(\eta_1)} \right)^{-1} \tag{17.2}$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$.

Theorem 15. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty}\frac{n-s-A}{\sqrt{A}}=\eta, \qquad -\infty<\eta<\infty; \tag{i}$$

$$\lim_{\lambda\to\infty}\sqrt{s}\left(1-\frac{\rho_a}{s}\right)=\beta, \qquad \beta=0 \tag{ii}$$

where all other parameters are fixed. Then

$$\lim_{\lambda\to\infty}P(W>0)=\left(1+\frac{\int_{-\infty}^{0}\Phi\left(\eta-t\sqrt{B}\right)d\Phi(t)}{\frac{1}{\sqrt{B}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right)}\right)^{-1} \tag{17.3}$$

where $\eta_1=\eta-\frac{\beta}{\sqrt{B}}$.

17.3 The Probability of Blocking

We can also calculate the probability of blocking. From the stationary probability, we will deduce the probability P_l that there are l customers in the system $(0\le l\le n)$. We will use the following relation:

$$\begin{aligned}
P_l &:= \sum_{\substack{i,i_1\ldots,i_K\ge 0\\ i+i_1\ldots+i_K=l}}\pi(i,i_1,\ldots i_K)=\sum_{i=0}^{l}\sum_{i_1+\ldots+i_K=l-i}\pi(i,i_1,\ldots,i_K)\\
&=\sum_{i=0}^{l}\sum_{i_1+\ldots+i_K=l-i}\pi_0\frac{1}{\nu(i)}(\rho_a)^i\prod_{j=1}^{K}\frac{1}{i_j!}(\rho_j)^{i_j}\\
&=\sum_{i=0}^{l}\pi_0\frac{1}{\nu(i)}(\rho_a)^i\frac{1}{(l-i)!}(A)^{l-i}\\
&=\pi_0\left(\sum_{i=0}^{s}\frac{1}{i!}(\rho_a)^i\frac{1}{(l-i)!}(A)^{l-i}+\sum_{i=s+1}^{n}\frac{1}{s!s^{i-s}}(\rho_a)^i\frac{1}{(l-i)!}(A)^{l-i}\right)
\end{aligned}$$

This phrase is exactly the same phrase that was investigated in Chapter 14. Thus, we can state the following theorem:

Theorem 16. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty}\frac{n-s-A}{\sqrt{A}}=\eta, \qquad -\infty<\eta<\infty; \tag{i}$$

$$\lim_{\lambda\to\infty}\sqrt{s}\left(1-\frac{\rho_a}{s}\right)=\beta, \qquad -\infty<\beta<\infty,\ \beta\neq 0 \tag{ii}$$

where all other parameters are fixed. Then

$$\lim_{\lambda\to\infty}\sqrt{s}P(block)=\frac{\nu\phi(\nu_1)\Phi(\nu_2)+\phi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{B}\right)d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)} \tag{17.4}$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$,$\nu_1 = \frac{\eta\sqrt{B^{-1}}+\beta}{\sqrt{1+B^{-1}}}$,$\nu_2 = \frac{\beta\sqrt{B^{-1}}-\eta}{\sqrt{1+B^{-1}}}$,$\nu = \frac{1}{\sqrt{1+B^{-1}}}$.

Theorem 17. *Let the variables* λ, s *and* n *tend to* ∞ *simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty} \frac{n-s-A}{\sqrt{A}} = \eta, \qquad -\infty < \eta < \infty; \tag{i}$$

$$\lim_{\lambda\to\infty} \sqrt{s}\left(1 - \frac{\rho_a}{s}\right) = \beta, \qquad \beta = 0 \tag{ii}$$

where all other parameters are fixed. Then

$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{0}\Phi\left(\eta - t\sqrt{B}\right)d\Phi(t) + \frac{1}{\sqrt{B}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta) + \phi(\eta)\right)} \tag{17.5}$$

where $\nu_1 = \frac{\eta}{\sqrt{1+B}}$, $\nu_2 = \frac{\eta}{\sqrt{1+B^{-1}}}$,$\nu = \frac{1}{\sqrt{1+B^{-1}}}$.

17.4 Expected Waiting Time

The last example stated here is for the approximation of E[W].

Theorem 18. *Let the variables* λ, s *and* n *tend to* ∞ *simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty} \frac{n-s-A}{\sqrt{A}} = \eta, \qquad -\infty < \eta < \infty; \tag{i}$$

$$\lim_{\lambda\to\infty} \sqrt{s}\left(1 - \frac{\rho_a}{s}\right) = \beta, \qquad -\infty < \beta < \infty,\ \beta \neq 0 \tag{ii}$$

where all other parameters are fixed. Then

$$\lim_{\lambda\to\infty} \sqrt{s}E[W] = \frac{\phi(\beta)\Phi(\eta) + \phi(\sqrt{\eta^2+\beta^2})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\left(B^{-1}\beta^2 - \eta\beta\sqrt{B^{-1}} - 1\right)}{\mu\beta^2\left(\int_{-\infty}^{\beta}\Phi\left(\eta + (\beta-t)\sqrt{B}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\right)}$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$.

Theorem 19. *Let the variables* λ, s *and* n *tend to* ∞ *simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty} \frac{n-s-A}{\sqrt{A}} = \eta, \qquad -\infty < \eta < \infty; \tag{i}$$

$$\lim_{\lambda\to\infty} \sqrt{s}\left(1 - \frac{\rho_a}{s}\right) = \beta, \qquad \beta = 0 \tag{ii}$$

where all other parameters are fixed. Then

$$\lim_{\lambda\to\infty} \sqrt{s}E[W] = \frac{1}{2\mu}\frac{B^{-1}\left((\eta^2+1)\Phi(\eta) + \eta\phi(\eta)\right)}{\sqrt{2\pi}\int_{-\infty}^{0}\Phi\left(\eta - t\sqrt{B}\right)d\Phi(t) + \sqrt{B^{-1}}\left(\eta\Phi(\eta) + \phi(\eta)\right)}$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$.

18 Defining Optimal Design

The goal of this chapter is to outline, very roughly, some methods for optimizing the system, using the approximations calculated in previous chapters. This optimization could be done in various ways. For example, from an economic point of view, one can find the optimal number of beds and nurses, so that the total cost is minimized while, at the same time, maintaining a predefined service level. Service level constraints could be on the delay probability $P(W > 0) < a$, the probability of waiting more than t units of time $P(W > t) < b$, or the probability of blocking $P(block) < c$. Any combination of these measures could be used as well.

Define C_n to be the annual bed costs due to space and maintenance, and C_s the annual nurse (servers) costs. Then our optimization problem can be:

$$
\begin{aligned}
&min_{n,s}\ C(n,s) = C_n n + C_s s \\
&s.t\ \ P(W > t) \leq b \\
&\quad\ P(block) \leq c \\
&0 \leq s \leq n.
\end{aligned}
$$

Another possibility could be to look at the situation as a revenue maximization problem. The hospital charges the insurance companies for each patient being hospitalized (under supervision on the necessity of the procedure). A patient that is being blocked is lost revenues to the system, as well as a threat to the hospital's reputation.

Define R to be the annual revenue due to bed occupancy, and $OC(n,s)$ the average bed occupancy level in a system with n beds and s nurses. This type of optimization problem can be formalized as:

$$
\begin{aligned}
&max_{n,s}\ R(n,s) = R \cdot OC(n,s) - C_n n - C_s s \\
&s.t\ \ P(W > t) \leq b \\
&\quad\ P(block) \leq c \\
&0 \leq s \leq n.
\end{aligned}
$$

One can also solve process-based optimization (control) problems, in support of real-time management, and we plan to pursue that in the future.

19 The Semi-Open Erlang-R Model

In this section, we examine the time-varying semi-open Erlang-R model. As explained in the Introduction, it could be viewed as an Erlang-R model with an additional upper bound on the number of customers in the system. Figure 43 depicts a graphical representation of our system. The number of customer (patients) is bounded by the number of spaces (beds) in the system, which is n. Customers that are blocked are thought of as being transfered to another system. In this case, the natural comparison is not to the Erlang-C (M/M/s) model, but to a loss system (M/M/s/$\tilde{n}$), where $\tilde{n}$ is properly chosen; specifically, $\tilde{n} = n - R_2$, where R_2 is the average number of content customers. We show that the steady-state distribution of the semi-open Erlang-R (more specifically, its Needy part) is different from that of a loss system. We use our results, from Section 17, to present the QED steady-state approximation of this model. We then show simulation results that demonstrate that, in the QED regime, the MOL approach works very well also in this case.

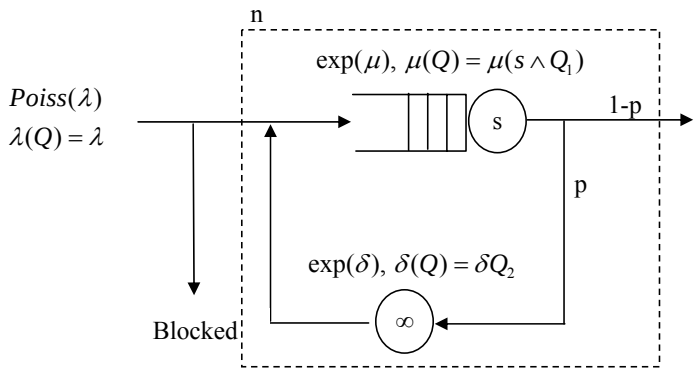


Figure 43: Semi-open Erlang-R model

We compare our model to a loss system in which the service time of a customer is the total work brought by that customer to the system, i.e., we sum up all the anticipated service times: a geometric sum of i.i.d. exponentials (which is thus exponential itself). See Figure 44 for a depiction of this loss model. Note that the average number of visits per customer (a) is more than 1.

Our comparison will start with a comparison of the steady-state probabilities of each model.

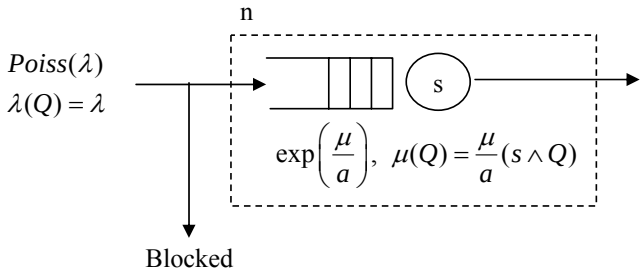


Figure 44: The loss model corresponding to a semi-open Erlang-R model

19.1 Steady State Comparison

19.1.1 Loss System (M/M/s/n) in Steady State

The steady-state distribution of a loss system is

$$\pi_k = \begin{cases} \frac{(S\rho)^i}{i!}\pi_0 & , \quad 0 \le i < S, \\ \frac{S^S\rho^i}{S!}\pi_0 & , \quad S \le i \le N, \end{cases}$$

where

$$\pi_0 = \left[\sum_{i=0}^{S-1} \frac{(S\rho)^i}{i!} + \sum_{i=S}^{N} \frac{S^S\rho^i}{S!}\right]^{-1},$$

and $\rho = \frac{\lambda}{S\mu}$. By the Arrival theorem [17], the probability of waiting (α) is

$$\alpha = \sum_{i=S}^{N} \pi_i^{N-1} = \frac{\sum_{i=S}^{N-1} \frac{S^S\rho^i}{S!}}{\sum_{i=0}^{s-1} \frac{(S\rho)^i}{i!} + \sum_{i=S}^{N-1} \frac{S^S\rho^i}{S!}}.$$

The probability of blocking is

$$P(block) = \pi_N = \frac{S^S\rho^N}{S!}\pi_0 = \frac{S^S\rho^N}{S!}\left[\sum_{i=0}^{S-1} \frac{(S\rho)^i}{i!} + \sum_{i=S}^{N} \frac{S^S\rho^i}{S!}\right]^{-1}.$$

QED Approximations of Performance Measures

The following theorem provides QED approximations of this loss system [49].

Theorem 20. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the following conditions:*

$$\lim_{\lambda\to\infty} \frac{n-s}{\sqrt{\frac{\lambda}{\mu}}} = \eta, \qquad 0 < \eta < \infty, \tag{i}$$

$$\lim_{\lambda\to\infty} \frac{s-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \qquad -\infty < \beta < \infty, \tag{ii}$$

where all other parameters are fixed. Then,

$$\lim_{\lambda\to\infty} P(W>0) = \begin{cases} \left(1+\frac{\beta\Phi(\beta)}{\phi(\beta)(1-e^{-\eta\beta})}.\right)^{-1} & , \quad \beta\neq 0, \\ \left(1+\frac{\sqrt{\pi}}{\eta\sqrt{2}}\right)^{-1} & , \quad \beta=0, \end{cases}$$

$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \begin{cases} \frac{\beta\phi(\beta)e^{-\eta\beta}}{\beta\Phi(\beta)+\phi(\beta)(1-e^{-\eta\beta})} & , \quad \beta\neq 0, \\ \frac{1}{\sqrt{\frac{\pi}{2}}+\eta} & , \quad \beta=0, \end{cases}$$

and,

$$\lim_{\lambda\to\infty} \sqrt{s}E[W] = \begin{cases} \frac{\frac{\phi(\beta)}{\mu}\left[\frac{1-e^{-\eta\beta}}{\beta}-\eta e^{-\eta\beta}\right]}{\beta\Phi(\beta)+\phi(\beta)(1-e^{-\eta\beta})} & , \quad \beta\neq 0, \\ \frac{1}{2\mu}\frac{\eta^2}{\eta+\sqrt{\frac{\pi}{2}}} & , \quad \beta=0. \end{cases}$$

Note that in our Multi-service loss system, μ is divided by a (the average number of returns to service, per customer).

19.1.2 Semi-Open Erlang-R in Steady State

Now we describe the steady-state distribution of the semi-open Erlang-R queue. Let us define

$$R_1 = \frac{\lambda}{(1-p)\mu},\ R_2 = \frac{p\lambda}{(1-p)\delta},\ \textit{and}\ B = \frac{R_1}{R_2} = \frac{\frac{\lambda}{(1-p)\mu}}{\frac{p\lambda}{(1-p)\delta}} = \frac{\delta}{p\mu}.$$

Then, the steady state distribution of this system is given by

$$\pi(i,j) = \begin{cases} \pi_0 \frac{1}{\nu(i)} (R_1)^i \frac{1}{j!} (R_2)^j & , \quad 0\leq i+j\leq n, \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Here $\nu(i)$ is defined as

$$\nu(i) := \begin{cases} i! & , \quad i\leq S, \\ S!S^{i-S} & , \quad i\geq S, \end{cases}$$

where π_0 is given by

$$\begin{aligned} \pi_0^{-1} &= \sum_{0\leq i+j\leq n} \frac{1}{\nu(i)} (R_1)^i \frac{1}{j!} (R_2)^j \\ &= \sum_{l=0}^{N} \frac{1}{l!} (R_1+R_2)^l + \sum_{l=S+1}^{n} \sum_{i=S+1}^{l} \left(\frac{1}{S!S^{i-S}} - \frac{1}{i!}\right) \frac{1}{(l-i)!} (R_1)^i (R_2)^{l-i}. \end{aligned}$$

The probability of waiting is given by

$$P_N(W>0) = \pi_0^{n-1} \sum_{s\leq i+j\leq n-1 | i\geq s} \frac{1}{s!s^{i-s}} (R_1)^i \frac{1}{j!} (R_2)^j,$$

where

$$\pi_0^{n-1} = \left[\sum_{0 \le i+j \le n-1} \left(\frac{1}{\nu(i)} (R_1)^i \frac{1}{j!} (R_2)^j \right) \right]^{-1} = \left[\sum_{i=0}^{n-1} \sum_{l=0}^{n-i-1} \left(\frac{1}{\nu(i)} (R_1)^i \frac{1}{l!} (R_2)^l \right) \right]^{-1}.$$

The probability of blocking is given by

$$P_N = \pi_0 \left(\frac{1}{N!} (R_1 + R_2)^N + \sum_{i=S+1}^{N} \left(\frac{1}{S! S^{i-S}} - \frac{1}{i!} \right) (R_1)^i \frac{1}{N-i!} (R_2)^{N-i} \right).$$

QED Approximations of Performance Measures

The following theorem states the QED approximations of our semi-open Erlang-R system.

Theorem 21. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the following conditions:*

$$\lim_{\lambda \to \infty} \frac{n - s - R_2}{\sqrt{R_2}} = \eta, \qquad -\infty < \eta < \infty, \tag{i}$$

$$\lim_{\lambda \to \infty} \sqrt{s} \left(1 - \frac{R_1}{s} \right) = \beta, \qquad -\infty < \beta < \infty, \tag{ii}$$

where all other parameters are fixed. Then,

$$\lim_{\lambda \to \infty} P(W > 0) = \begin{cases} \left(1 + \frac{\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{B}\right) d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2} \Phi(\eta_1)} \right)^{-1} & , \quad \beta \neq 0, \\ \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{B}\right) d\Phi(t)}{\frac{1}{\sqrt{B}} \frac{1}{\sqrt{2\pi}} (\eta \Phi(\eta) + \phi(\eta))} \right)^{-1} & , \quad \beta = 0, \end{cases} \tag{19.1}$$

$$\lim_{\lambda \to \infty} \sqrt{s} P(block) = \begin{cases} \frac{\nu \phi(\nu_1) \Phi(\nu_2) + \phi(\sqrt{\eta^2+\beta^2}) e^{\frac{\eta_1^2}{2}} \Phi(\eta_1)}{\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{B}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2} \Phi(\eta_1)} & , \quad \beta \neq 0, \\ \frac{\nu \phi(\nu_1) \Phi(\nu_2) + \frac{1}{\sqrt{2\pi}} \Phi(\eta)}{\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{B}\right) d\Phi(t) + \frac{1}{\sqrt{B}} \frac{1}{\sqrt{2\pi}} (\eta \Phi(\eta) + \phi(\eta))} & , \quad \beta = 0, \end{cases} \tag{19.2}$$

and,

$$\lim_{\lambda \to \infty} \sqrt{s} E[W] = \begin{cases} \frac{\phi(\beta)\Phi(\eta) + \phi(\sqrt{\eta^2+\beta^2}) e^{\frac{1}{2}\eta_1^2} \Phi(\eta_1) \left(B^{-1}\beta^2 - \eta\beta\sqrt{B^{-1}} - 1 \right)}{\mu\beta^2 \left(\int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{B}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\eta_1^2} \Phi(\eta_1) \right)} & , \quad \beta \neq 0, \\ \frac{1}{2\mu} \frac{B^{-1}\left((\eta^2+1)\Phi(\eta) + \eta\phi(\eta) \right)}{\sqrt{2\pi} \int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{B}\right) d\Phi(t) + \sqrt{B^{-1}} (\eta \Phi(\eta) + \phi(\eta))} & , \quad \beta = 0. \end{cases}$$

Here $B = \frac{R_1}{R_2} = \frac{\delta}{p\mu}$, $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$, $\nu = \frac{1}{\sqrt{1+B^{-1}}}$, $\nu_1 = \frac{\eta\sqrt{B^{-1}} + \beta}{\sqrt{1+B^{-1}}}$, $\nu_2 = \frac{\beta\sqrt{B^{-1}} - \eta}{\sqrt{1+B^{-1}}}$.

19.1.3 Comparing Steady-State Measures

When comparing exact system measures, we observe that the steady-state measures of the loss system are not equal to the ones of the semi-open Erlang-R model. Indeed, there is no simple relation between them, since the meaning of n in each model is different. In the semi-open Erlang-R model n is the number of Needy and Content customers together, while in the loss system it represents only the number of Needy customers. As a result, the definition and meaning of η in each system is different. Therefore, we expect that when setting the same number of servers s and customers n in both systems, the loss system will always underestimate $P(block)$ and overestimate $P(W > 0)$.

To understand the differences between the models, we compared the steady-state service measures of a semi-open Erlang-R and a corresponding loss system. For a correct comparison, we compare models with the same number of servers s (or β), while adjusting the number of beds n. The parameter n in the semi-open Erlang-R model is larger than that of the loss model - denote by latter by $\tilde{n}$. The difference $n - \tilde{n}$ is taken to be the average number of Content customers R_2.

The performance measures of both models are a function of the decision variables s and n, and the offered loads. In fact, performance of the semi-open Erlang-R, and similarly to the open one, depends on the ratio between the stations' offered load: $B = \frac{R_1}{R_2}$, as in Theorem 21. We refer to B as the offered load ratio.

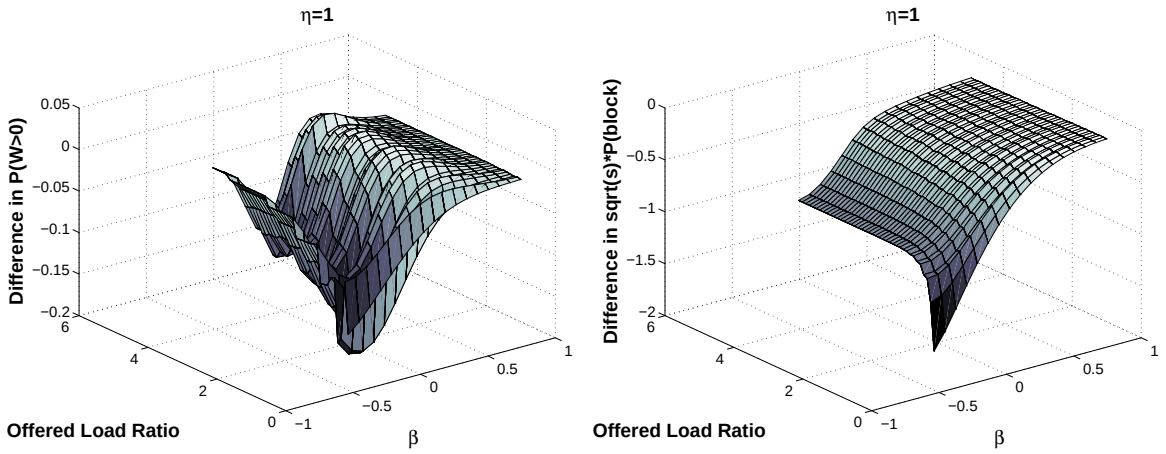


Figure 45: Difference between service measures as a function of Offered Load Ratio and s

Figure 45 shows the difference between the service measures of the two models as a function of B and β. We observe that the difference in $P(W > 0)$ is largest when $\beta \in [-1, 1]$, which is the QED regime. The difference in $P(block)$ seems to grow as β and B decrease. The difference in $P(W > 0)$

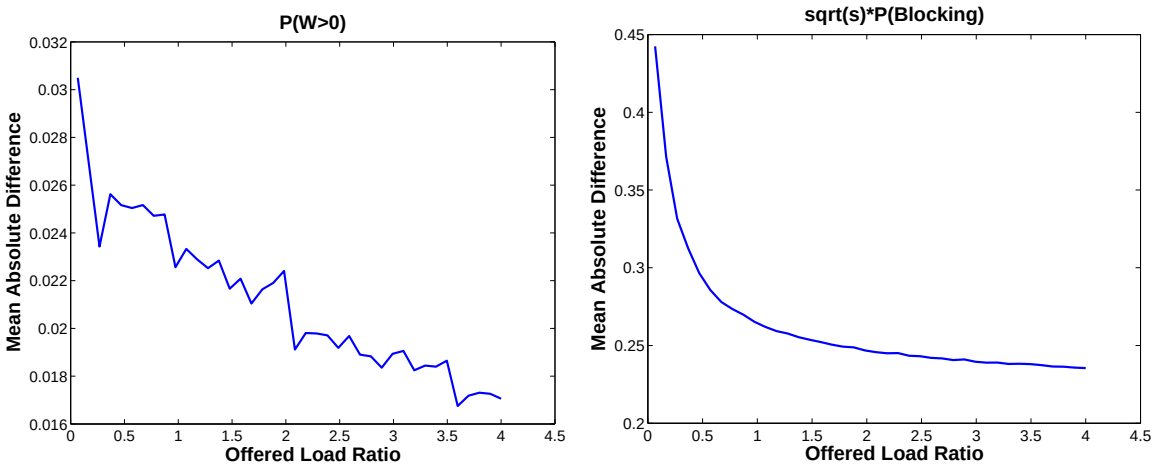


Figure 46: Mean difference between service measures as a function of Offered Load Ratio

and $\sqrt{s} * P(block)$ can be up to 20%. In order to understand more precisely the influence of the offered load ratio, we also calculated the mean absolute difference between the service measures of the two models. Figure 46 shows this mean difference as a function of B. We observe that the difference in $P(W > 0)$ and $P(block)$ is decreasing when B grows. We conclude that one must take into account reentering customers especially when the difference is significant i.e. for B smaller than 1.

We calculated the offered-load ratio of four medical units: an Internal Ward, Emergency Ward, ICU, and Oncology Ward. The corresponding offered-load ratios are 0.1, 0.4, 2, 0.2. Thus, in all cases but the ICU, when using a loss model for planning, without considering the Re-entering effect, the measures $P(W > 0)$ and $P(block)$ are both underestimated, and the planner will therefore under-staff doctors (or nurses) and beds in the system. These under-estimations prevail uniformly over most values of β. Such under-staffing relative to the number of beds results in long waits for service and personnel burnout due to high workload.

19.2 Staffing Semi-Open Erlang-R with Time-Varying Arrivals

In this section, we verify the usefulness of the MOL approximation for our semi-open queueing network. We use the following staffing rule: Set s (the number of servers) by

$$s(t) = R_1(t) + \beta\sqrt{R_1(t)},$$

and then set n (the number of beds) via

$$n(t) = s(t) + R_2(t) + \eta\sqrt{R_2(t)}.$$

Here $R_1(t)$ and $R_2(t)$ are determined by the fluid ODE of the regular Erlang-R system (5.2), and β and η according to the steady-state QED approximations (19.1) and (19.2), respectively.

To analyze our approach, we use a simulation with the parameters of Case Study 1. Figure 47 presents steady-state QED approximations for the probability of waiting and the blocking probability. Figure 48 presents simulation results of the system when staffing according to the square-root formula with $\eta = 1$, and various values of β. Figure 49 shows a comparison between the theoretical steady-state QED approximations and our simulations results.

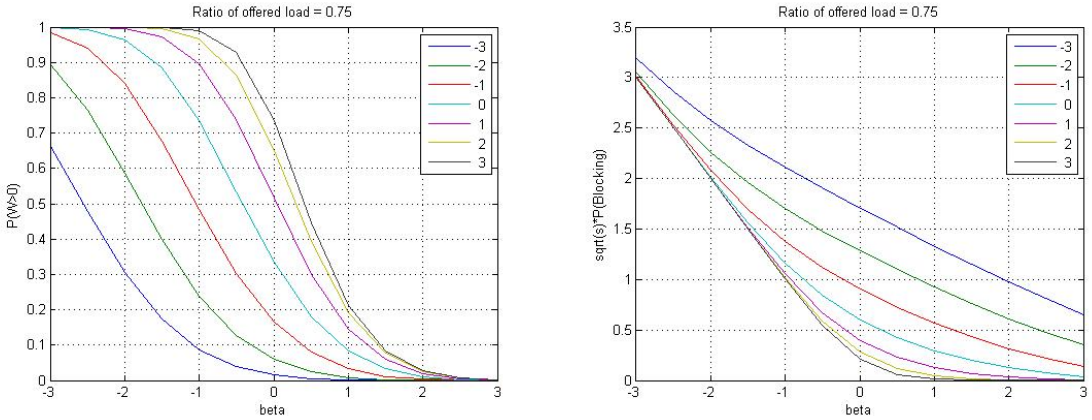
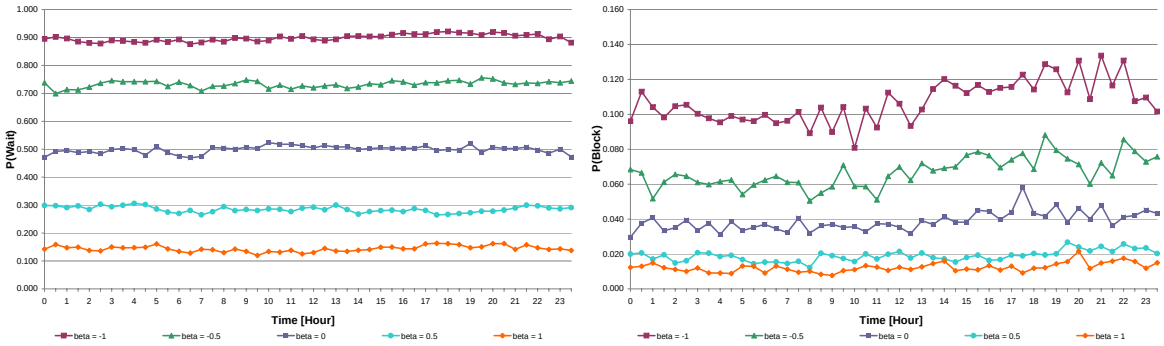


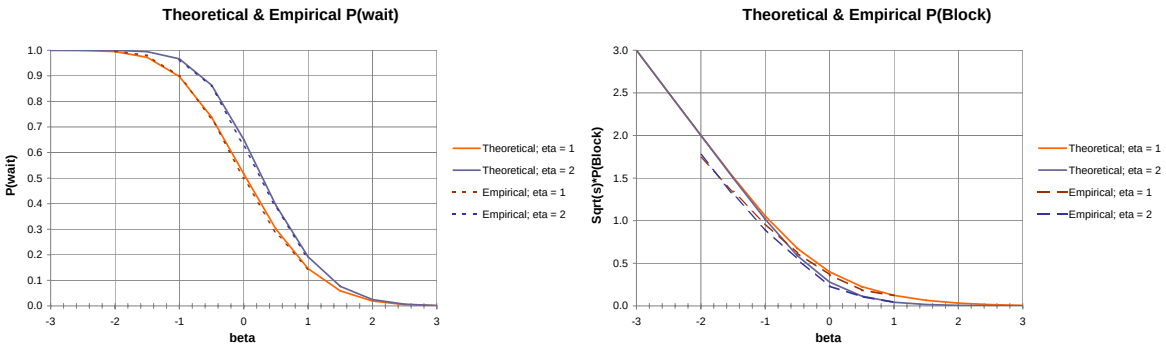
Figure 47: Steady state $P(W > 0)$ and P(block) as a function of β and η, for semi-open Erlang-R

We observe that the probability of waiting is stable and its average fits steady-state values. When considering the probability of blocking, we note good results only for β values that exceeds 0. This result is consistent with the observations showed in Section 15, that the $P(W > 0)$ approximation is stable over all operational regimes, but the $P(block)$ approximation is stable only in the QED regime.



$P(W > 0)$ P(block)

Figure 48: $P(W > 0)$ and P(block) changing over time, for semi-open Erlang-R



$P(W > 0)$ P(block)

Figure 49: Average $P(W > 0)$ and P(block) as a function of β, for semi-open Erlang-R

20 Managerial Insights

In this section, we analyze the behavior of some system measures in the QED operational regime. We start with describing each measure separately.

20.1 Behavior of the Probability of Waiting

Theorems 8 and 9 are the indication that the policy described in Equation (13.1) is in fact a QED policy; indeed, the probability of waiting converges to a limit that is strictly between 0 and 1. Notice that the waiting probability is a function of only three parameters: β and η, which are decision variables, and the offered load ratio B, which captures the physics of the system. The offered load ratio is the ratio between the offered load of service station and the total offered load of the non-service stations.

20.2 Behavior of the Probability of Blocking

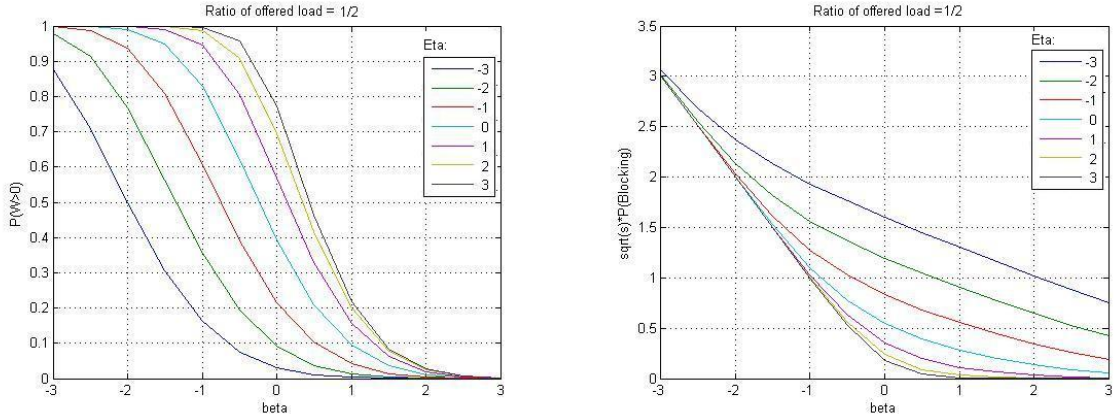
Theorems 12 and 13 demonstrate that the probability of blocking is in the order of $\frac{1}{\sqrt{s}}$. For example, assume that the offered load ratio is 0.5 and the system in large (100 servers). Using figure 50, we observe that, by choosing the pair $\eta = 1$ and $\beta = 0.2$, we actually aim at a probability of getting served immediately to be 50%. At the same time, the probability of getting immediately a bed is 95%. Thus, our QED policy gives more importance to blocking than to waiting.

20.3 Behavior of the Expected Waiting Time for a Nurse

Theorems 10 and 11 show that the expected waiting time is in the order of $\frac{1}{\sqrt{s}}$. Note that the expected waiting time divided by the expected service time is also of that same order, which means that for large systems the wait time is one order of magnitude less then the service time.

20.4 Influence of β and η

Since the physics of the QED system is driven by only three parameters, one can describe their influence in a very simple way. For example, if the offered load ratio is 0.5, the two graphs in Figure 50 tell us all we need to know about the probability of waiting and blocking, and illustrate the influence of the decision variable β and η on these systems' performance measures. In fact, the right graph shows the probability of blocking multiplied by $\sqrt{s}$, and thus, those graphs fit any unit size, 30 or 300 alike.



P(W>0) $\sqrt{s}$P(block)

Figure 50: Demonstration of the influence of β and η on $P(W > 0)$ and P(block)

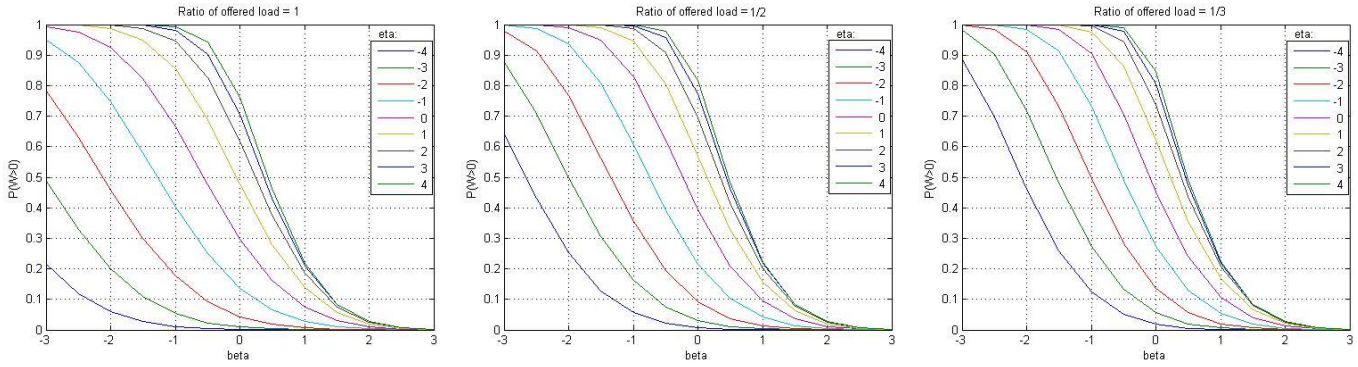
One can observe that, as β grows, which means that the number of nurses grows, both the probability of waiting and the blocking probability tend to zero. When there are fewer nurses (β decreases), the probability of waiting approaches 1, and the blocking probability, multiplied by $\sqrt{s}$, approaches β. One also observes the sensitivity of the system measures to β and η, and that one need not consider the range $(-\infty, \infty)$, but a much smaller range such as $\beta \in (-3, 3)$ suffices. If we fix the number of nurses (β), we see that, as the number of beds (η) increases, the probability of waiting approaches 1, since there are more patients in the system, and the blocking probability approaches 0. Again, there is almost no difference between $\eta = 2$ and $\eta = 3$, so there is an effective limit to the beds needed, and the range of η's that must be considered is relatively small (also between -3 and 3). In addition, we saw in Section 15 that the QED regime is achieved in even narrower range of parameters ($\eta > -1$ and $\beta > -0.5$). Hence, we conclude that the relevant parameter ranges for the QED regime are $\beta \in (-0.5, 3)$ and $\eta \in (-1, 3)$.

Formally, our observation is that $P(W > 0)$ is a decreasing function with respect to β and η, while $\sqrt{S}P(block)$ is a decreasing function with respect to β and an increasing function with respect to η.

20.5 Influence of the Offered Load Ratio

The graphs in Figure 51 demonstrate the influence of the offered load ratio (B) on the probability of waiting. Recall that the offered load ratio is the ratio between the offered load in the service station and the total offered load of the non-service stations. Our observation is that $P(W > 0)$ is a decreasing function with respect to B. As the offered load ratio decreases, the probability to wait

increases since the ratio between the number of nurses and the number of beds decreases. Thus, there are more patients per nurse.



(a) (b) (c)

Figure 51: Comparison between different ratios - The influence on $P(W > 0)$

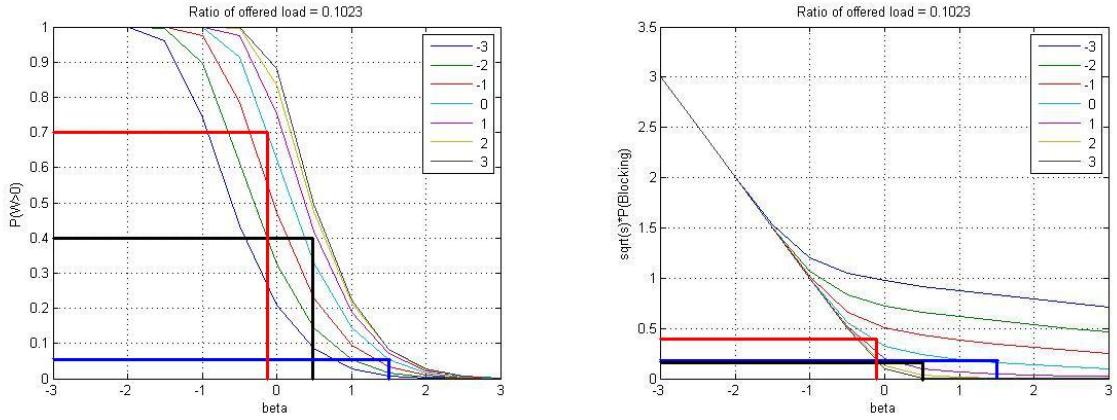
Note that the offered load ratio can be interpreted as a measure of the service intensity required during hospitalization; it is a natural measure to differentiate between patient types. More specifically, the fact that wards differ in patients' clinical conditions can be translated to a difference in the patients' cycle time of needy-dormant states which, in turn, can be analyzed by considering different offered load ratios (B). As B grows, patients require more nursing in each cycle. For example, in IWs the offered load ratio is around 0.1 (see Section 20.6 for en explanation), while in ICUs, where patients require more intensive care over longer periods, the offered load ratio is around 2.[3]

20.6 Demonstrating Three Operational Decisions

To illustrate staffing and allocation decisions, we use data originated from two articles: Lundgren and Segesten [52] and Green and Yankovic [36]. Green and Yankovic describe a medical unit that has 42 beds, with average occupancy level of 78%, and Average Length of Stay (ALOS) of 4.3 days. Lundgren and Segesten studied nurses' service times in a medical-surgical ward. They found that the average service time in their unit was 15.3 minutes per service, and that the average demand rate for each patient is 0.38 requests per hour. Therfore, we take an average service time of 15 minutes and assume that there are 0.4 requests per hour from each patient. Fitting this data to our model results in the following parameters values: $\lambda = 0.32, \mu = 4, \delta = 0.4, \gamma = 4, p = 0.975$ and the

[3]ICU's offered load ratio (B) was calculated based on the following data: ALOS is 9.5 days, average service time is 10 minutes and patients require 4 treatments per hour. Thus, $\mu = \gamma = 6, \delta = 12$ and $p = 0.99887$ which results in $B = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)} = 2$.

offered load ratio is then approximatly 0.1.



P(W>0) $\sqrt{s}$P(block)

Figure 52: $P(W > 0)$ and P(block) for ratio 0.1

Figure 52 shows the appropriate graphs for that offered load ratio, and an illustration of three available policies. The first pair (black) with $\beta = 0.5$ and $\eta = 0.5$, is QED balanced. It has 38 beds with 4 nurses. This policy combines low probability of blocking - 0.09 and a reasonable probability of waiting - 0.45. The second pair (blue) is more of a QD policy; we chose $\beta = 0.5$ and $\eta = -0.05$, which amounts to 37 beds and 6 nurses. In this setting, both waiting and blocking are low; $P(W > 0) = 0.09$ and $P(block) = 0.08$. The third example (red) is more of a ED, we chose negative $\beta = -0.1$ and $\eta = -0.05$ (which means a system with 34 beds and only 3 nurses). In this policy, the waiting and the blocking are relatively high; $P(block) = 0.19$ and $P(W > 0) = 0.72$.

20.7 Time Varying Environments

In section 19, we compared our semi-open Reentrant model to a loss model that does not explicitly account for Reentring customers. We saw that, even in steady state, there is importance for using our semi-open Erlang-R model, and preferring it over the simpler models. We saw how simple models under- or over- estimate system performance. Note that in medical systems, such as Emergency Wards, the offered-load ratio is small (around 0.4), and thus when using a loss model for planning, without considering the Re-entering effect, the measure of $P(W > 0)$ is underestimated, and the planner will therefore under-staff doctors (or nurses) in the system, and at the same time the measure of $P(block)$ is overestimated, which will cause the planner to recommend too many beds. This will result in long waits for service and personnel burnout, due to high work-load. These under- and over-estimation prevail uniformly over all values of β.

In addition, we saw that, in time varying environments, for a large enough system in the QED regime ($\beta >> -0.5$), the MOL approach stabilizes both $P(W > 0)$ and P(block). Consequently, any pre-specified QED service level can be achieved stably over time. We also observe that there is a small distortion in the P(block) results, demonstrating that the probability of blocking will actually be less then predicted.

21 Conclusions and Future Research

In this part, we have developed a model for medical units that incorporates two decision levels into one. We are able to use this model to set both staffing levels inside the Medical Unit, as well as design the size of the unit itself. We proposed an appropriate staffing procedure that carefully balances the system in the QED regime, in a way that stabilizes over time both utilization levels and service-level performance measures, such as the probability of waiting to a nurse and the probability of being blocked when seeking an available bed in the ward. We developed QED approximations that provide the expected service level under this procedure. We validated our approximations in various settings of system size and parameters, and showed that it works well both in large and small systems, as well as in realistic cases of an Internal Ward. We used QED approximations to better understand the dynamic of our system, and the influence of the various parameters.

We then developed a generalization of our QED approximations, to a wider set of networks, and used it for the analysis of the semi-open Erlang-R model. We showed that, in time-varying environments, one can stabilize both $P(W > 0)$ and P(block) over time, using a variation of the MOL staffing procedure. When comparing the semi-open Erlang-R to the non-reentrant system (M/M/s/n), we demonstrated analytically that using the simpler model in Re-entrant situations is detrimental, both in steady-state and in time varying environments.

Our models have natural extensions, such as multi-classes of patients or nurses, additional phases of clinical treatment, adding doctors (most likely working in the ED regime, in parallel to nurses in the QED regime), and random parameters. We have given thought to some possible such extensions, and we describe some of them in greater detail in Part IV.

Part III

Empirical Analysis of Patients-Flow Data

22 Introduction

In this part of the research, we provide details of an extensive empirical analysis that we conducted on patient-flow data at one of the largest hospitals in Israel. This hospital has approximately 1000 beds and 45 medical units. It provides service to about 75,000 patients annually. The hospital provided us with four years of data pertaining to all of their patients' transfers throughout the medical wards. We cleaned this data and used it to better understand the implication of various operational decisions. The empirical analysis we provide here is mainly concerned with the Hospitalization Wards, and focuses mostly on IWs. Additional analysis of this data can be found in Marmor [56] and Tseytlin [71].

We start, in Section 23, with empirical analysis of arrivals to IWs. We show how the arrival rate changes over scales of hours, days and months, and during special times such as the 2006 Israel-Lebanon War. We compare the IWs' arrival rate patterns to the ED's arrival rate, demonstrating a time-delay between the two, which are due to the LOS in the EW. Then, in Section 26, we analyze the LOS distribution in the IWs over two time-scales: days and hours. We provide operational explanations for the unique shape of the daily LOS histogram. We compare the four IWs, and show that all their LOS distributions have the same shape, although the parameters of these distributions are different. For example, the expectation of the LOS varies among wards. We verified the operational reasons for those differences by interviewing the medical staff and management of the wards. We also consider how one can simulate such special LOS distributions. In addition, a more detailed analysis was performed to investigate the dependence between LOS and the load in the Ward.

In Section 25.1, we demonstrate how arrival, departure and LOS become integrated in the WIP measure, and show how the number of patients in the ward (WIP) changes over time. We then discuss its influence on the workload of nurses.

In Section 24, we investigate the frequency of blocking at the IWs. We show that the blocking of the IWs experienced significant growth over the years. We show that this is to be expected when considering trends in the arrival rate and the reduction of available beds. We also provide other operational explanations that were deduced from the interviews.

Lastly, in Section 27, we compare other medical wards (e.g. Oncology) to the IW. We find several

differences, such as in the probability to return within three month, and the LOS distribution. We discuss medical and operational reasons for those differences.

22.1 Data Description

This documentation describes patient-level data at the Hospital. The data was recorded over the following periods: 1/1/2004 - 1/12/2008. There is a record (line in the file) for each patient's transfer in the hospital. The following are the fields for each record:

- Key - a unique number which identifies the ID of each patient.
- AdmissionNo - Patients during a particular visit in the hospital are identified by a serial number. Thus, the Key remains constant across visits, while the AdmissionNo changes by visit.
- FALAR (ED-Ward) - A code of the location type. ("3 (ED)" - Emergency Department, "1 (Ward)" - A medical ward).
- BEWTY (MoveType) - A code of the transfer transaction type. ("1 (Enter)" - enter the hospital, "2 (Exit)" - exit the hospital, "3 (Move)" - transfer inside the hospital between wards, "6 (Exit V)" - patient exits the hospital for a "vacation", "7 (Return V)" - patient returns to the hospital from "vacation").
- D-BWIDT - Patient's arrival date to the ward.
- D-BWEDT - Patient's exit date from the ward.
- BWIZT - Patient's arrival time to the ward.
- BWEZT - Patient's exit time from the ward.
- ORGPF - Nursing ward code (the physical ward where the patient is hospitalized).
- ORGFA - Medical ward code (the ward that is actually in charge of the patient, which could differ from the ward where the patient physically resides).

23 Arrivals to the Internal Wards

In this section, we describe the trends and patterns of arrival rates to the Internal Wards, in various resolutions. The analysis provided here is only of the average arrival rate, and does not describe the stochastic variability of the arrival process. The analysis is carried out in three time resolutions: yearly, weekly, and daily. Each time-scale serves different hierarchy level in the hospitals' work-force planning process, as described in Section 10.3.3.

Figure 53 describes the arrival rates per month from March 2004 to October 2008. Arrivals vary between 900 and 1,150 patients per month. There is general increasing trend in the number of patients of 1.7% per year. We also observe a special period between July and August 2006, in which the arrival rate is much smaller, around 850-900 patients per month. This is the effect of the 2006 war in the north of Israel.

In order to identify monthly patterns, we calculated the average monthly arrival rates. This data is presented in Figure 54, with minimum and maximum values for each month (the data from 7-8/2006 were excluded). We see that January is the month with the highest arrival rate during the year, with an average arrival rate of 1,150 patients per month. This fact is in agreements with the common winter overcrowding of IWs. From January to April, the rate deceases, until it reaches a minimum rate of 950 patients per month. Then the rate slightly increases and stabilizes on 1,050 patients per month.

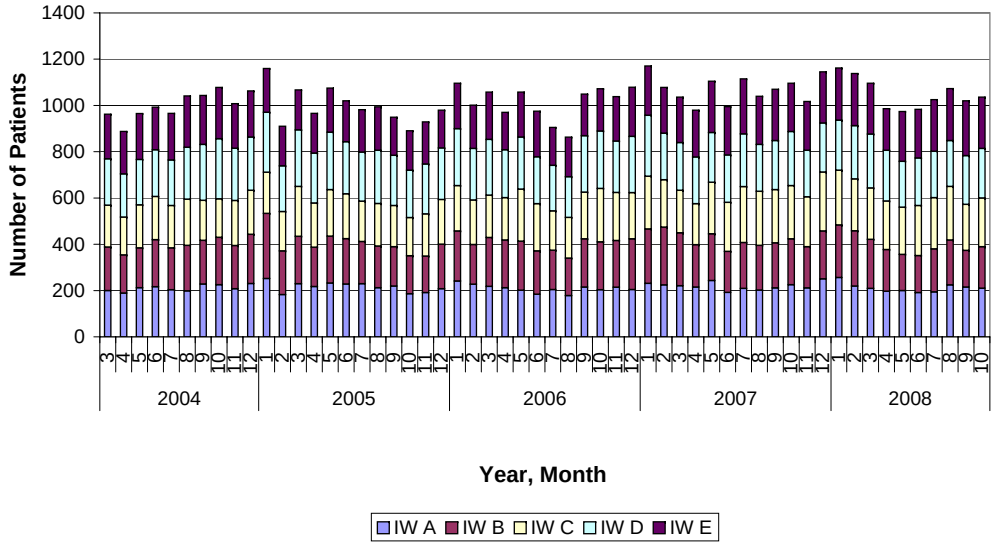


Figure 53: Arrival by year and month

When considering weekly patterns of arrival rates, we identify the following pattern (see Figure

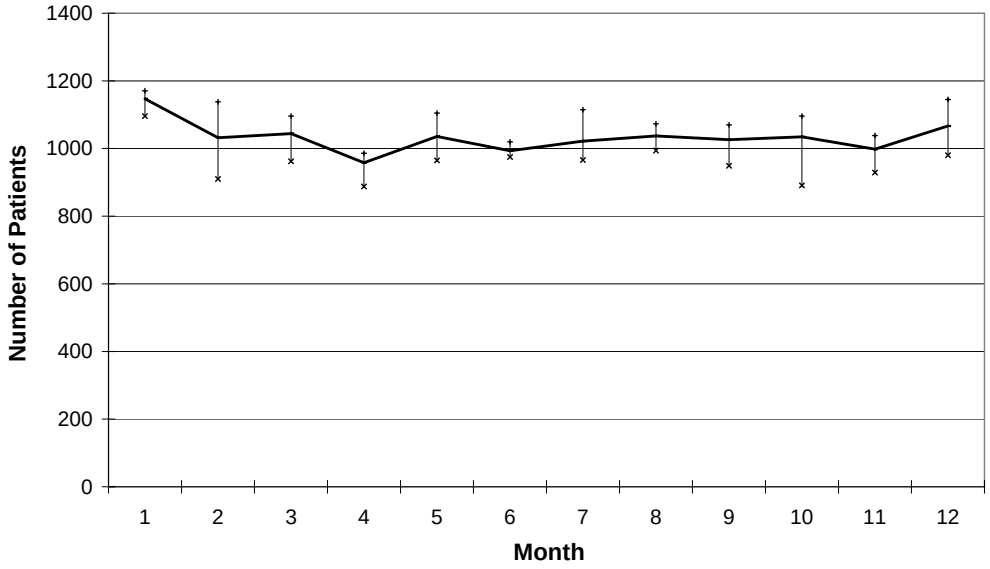


Figure 54: Arrival by month

55): The rate is highest at the beginning of the week (Sunday-Tuesday) and lowest during the weekend (Friday-Saturday). This pattern is valid for regular patients but not for Ventilated (V) and ICU (Intensive Care Unit) patients as can be seen in Figure 56. The arrival rate of the severe patients is constant over the days of week.

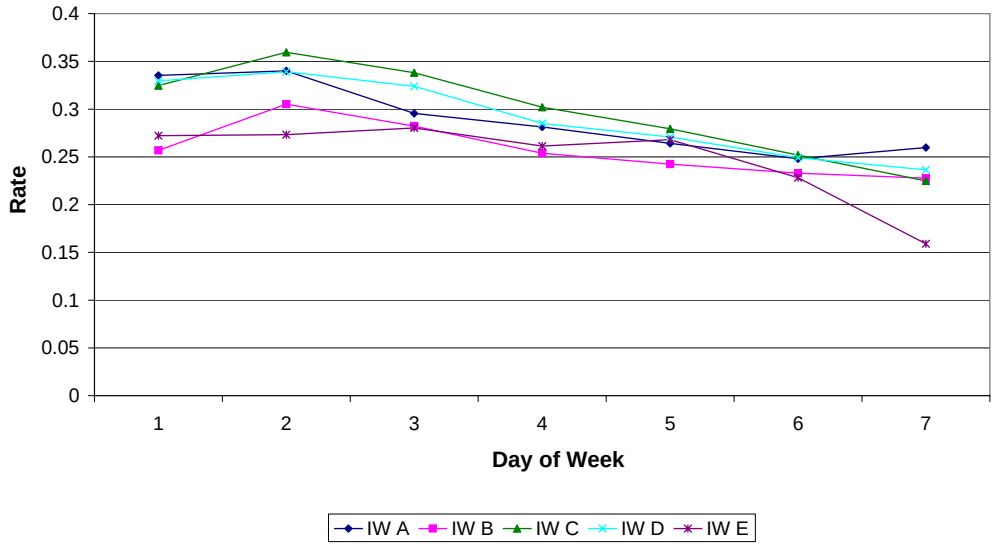


Figure 55: Arrival by day of week

When considering daily patterns, the arrivals to IWs are determined by the arrival process to the EW. We know that patients arrive at the hospital according to a time-dependent arrival rate [56]. Most patients arrive to the ED during the day and move to the IWs in the late afternoon.

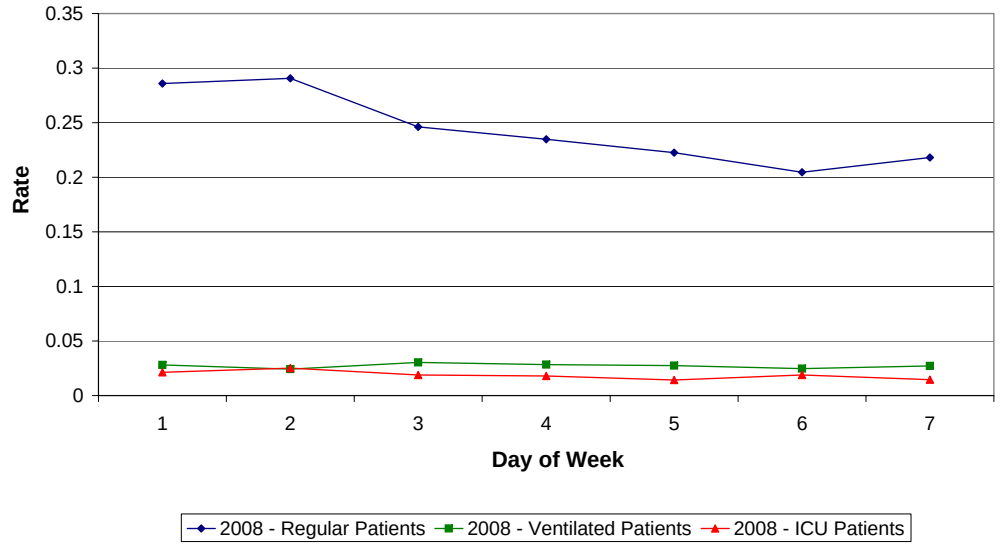


Figure 56: Arrival by day of week and patient type (Regular, V, ICU)

Figure 57 shows these daily patterns together. The process of admissions to the wards (transferals from the ED or other wards) has a pattern similar to the general arrivals but it is shifted in time. The time-lag between them is due to the LOS in the EW. Figures 58 and 59 show the weakly and daily arrival rate patterns of all IWs. We observe that most patients arrive to the IW between 12:00 - 02:00; late night and morning hours arrival rates are much lower. We also observe that weekends behavior is close in shape to weekdays behavior, though the overall volume during weekends is lower.

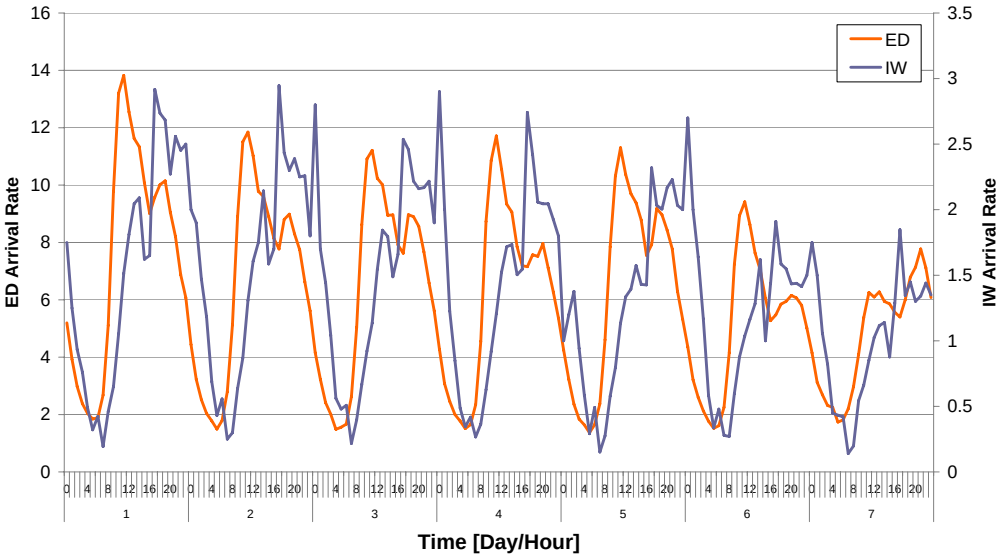


Figure 57: Arrival to EW and IW by day of week and hour

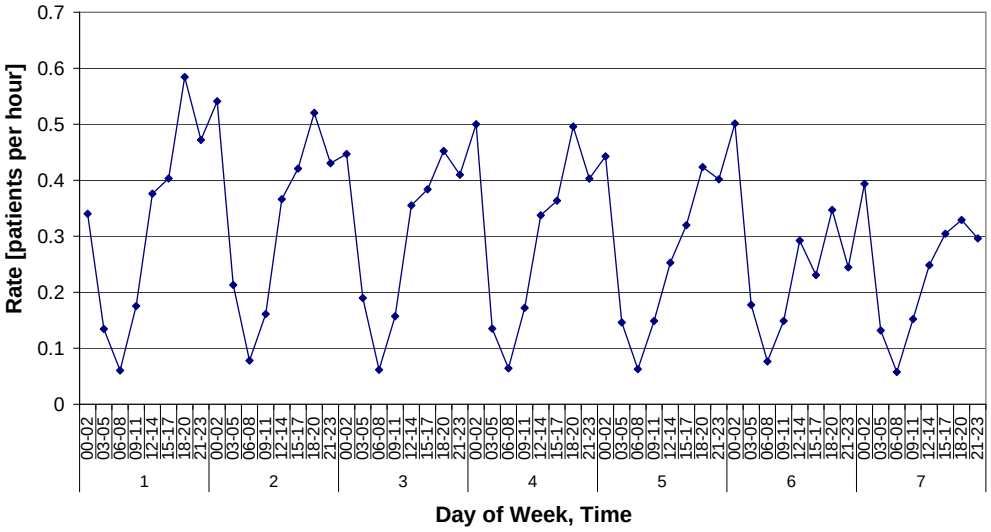


Figure 58: Arrival rate by day of week and hour

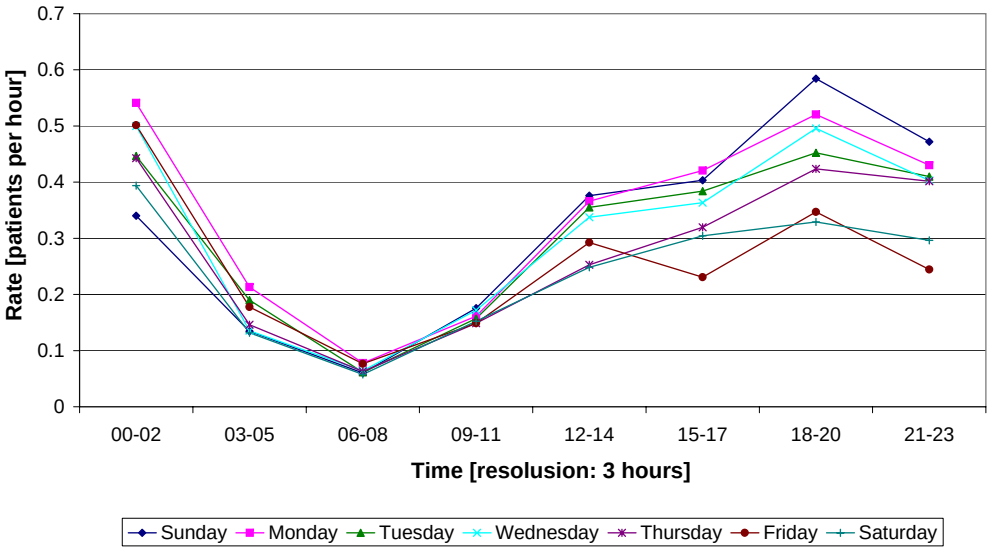


Figure 59: Arrival rate by hour

24 Blocking at the Internal Wards

The number of beds in each ward is limited. Hence, there are cases in which a patient is referred to a specific ward but there are no available beds for hospitalization in that ward. In such cases, the patients are *blocked* in the ED and transfered for hospitalization in other wards. The doctors of the originally-assigned ward will be in charge of the medical needs of the patient, but nursing treatment will be the responsibility of the actual hospital ward where the patient is located. Therefore, there is a distinction in the data between the *Nursing-ward* code, which is the code for the physical ward where the patient is hospitalized, vs. the *Medical-ward* code, which is the code for the ward that is actually in charge of the patient. Using this difference, one can estimate and analyze the frequency of blocking. There is also a less significant phenomena where, under these circumstances, another patient is actually transferred (or released) from the referred-to ward in order to vacate a bed. We do not estimate these incidents.

The number of beds in the IWs changed over the years. Figure 60 shows the number of beds available in all the IWs during the period 2004-2008. We observe that, in general, there is a reduction in the number of beds. At the beginning of 2007, there was a period during which a new "half-ward" with 20 beds was opened on a temporary basis. We also observed in Section 23 that the arrival rate increased over the years. Hence, we anticipate that blocking incidents will increase.

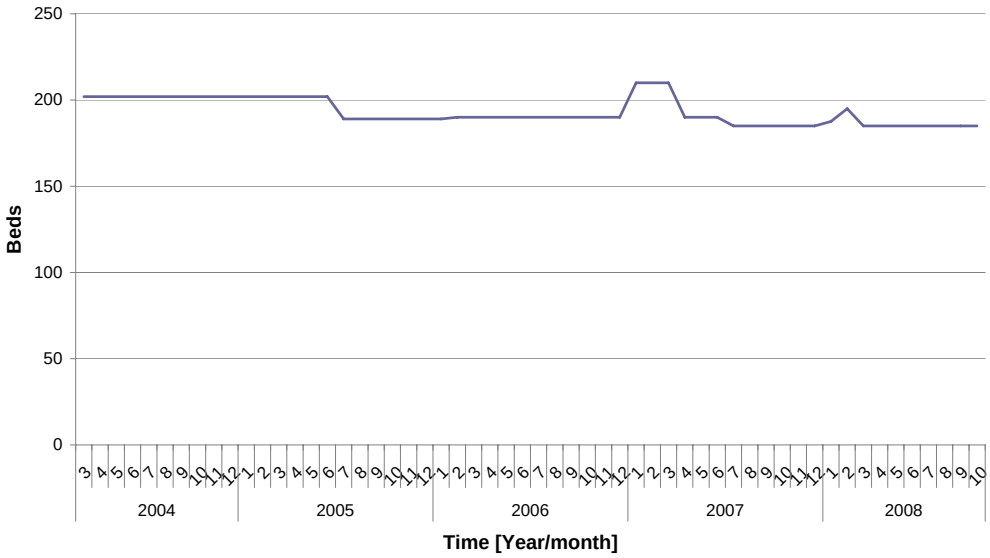


Figure 60: Number of beds in IWs by year and month (2004-2008)

Figure 61 describe blocking incidents that happened during the studied period. In general, the blocking percentage is around 3% (average of 30 incidents per month, when the arrival rate is

around 1000 patients per month). Figure 61 shows that, until 2007, blocking occurred mainly in January which is the month with the highest arrival rate. But from year 2007, blocking became a routine event. From 2007, almost 20% of the patients in IW E were blocked every day (see Figure 62). Some of the change is explained by the incease in arrivals and decrease in hospital beds (64). When examined in more details, we find that this is not the whole story since peaks in blocking do not match peaks in arrivals during that period. We have investigated the change we saw by intervieiwing hospital managment, and discovered that, at the beginning of 2007, the hospital started renovating IW D, and decreased the number of hospital beds in this ward. Therefore, the hospital decided to dedicate some of the beds in Ophthalmology Wards to IW patients; in fact, the Ophthalmology Ward became a branch of IW E. From that point on, patients were transferred to the Ophthalmology Wards on a regular basis, not only when blocked. Hence, in this graph, we see two seperate phenomena that represent two blocking policies: until 2007, blocking only when necessary, during overcrowding periods; and after 2007, load balancing between IWs and a new "buffer" ward.

Figure 63 illustrates the blocking phenomena on Day Of Week (DOW) basis, along with the number of patients in the ward, during the period 2007-10/2008. This is an interesting graph since it shows that blocking is stable during the week although arrival rates and the number of patients during the weekend are much smaller. This might be also a consequence of routinely sending patients to the buffer ward, balancing that continue also during weekends, as there are fewer personnel in the wards during weekends.

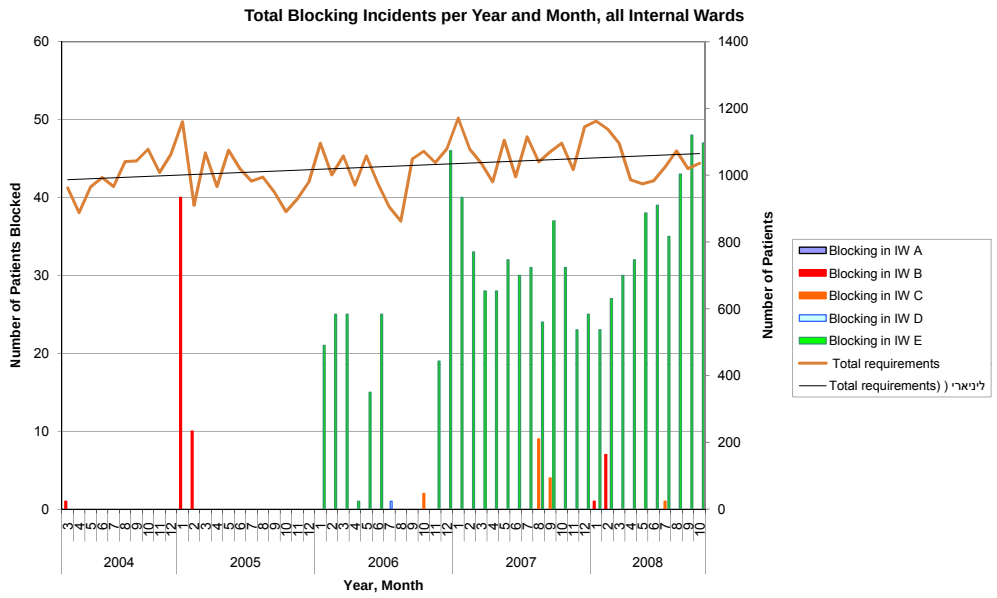


Figure 61: Number of patients blocked by year, month, and ward (2004-10/2008)

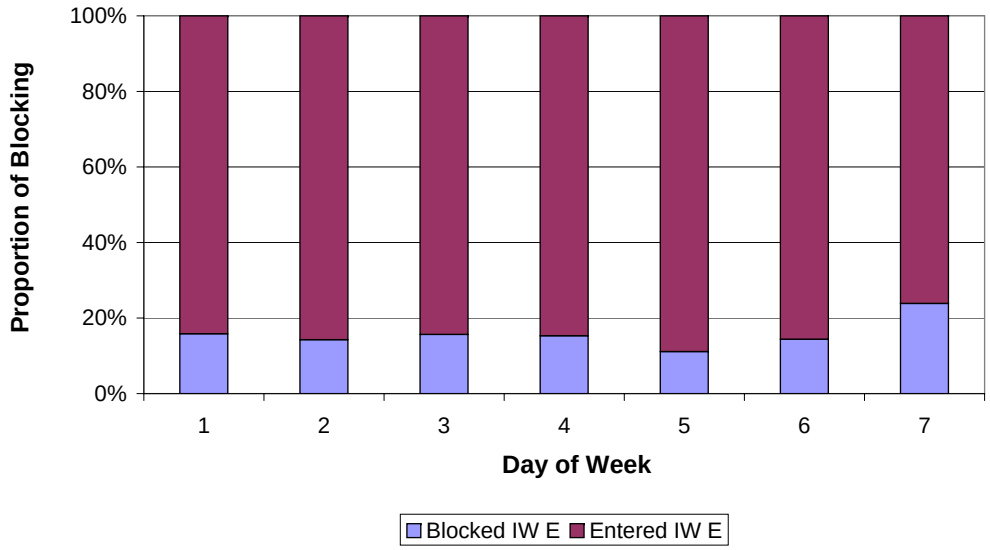


Figure 62: Percent of patients blocked by day in Ward E (2007-10/2008)

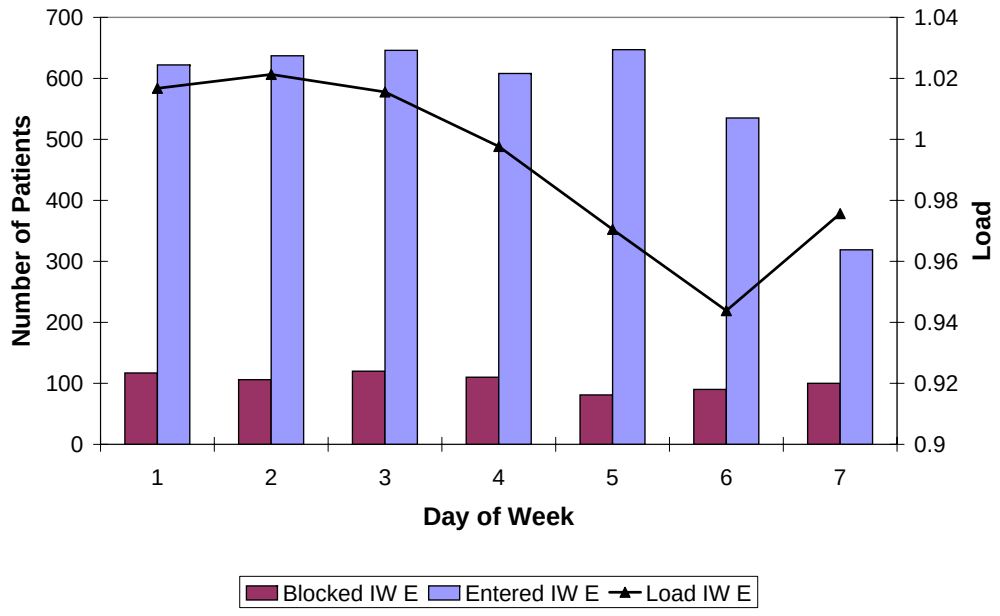


Figure 63: Number of patients blocked and patients load by day in Ward E (2007-10/2008)

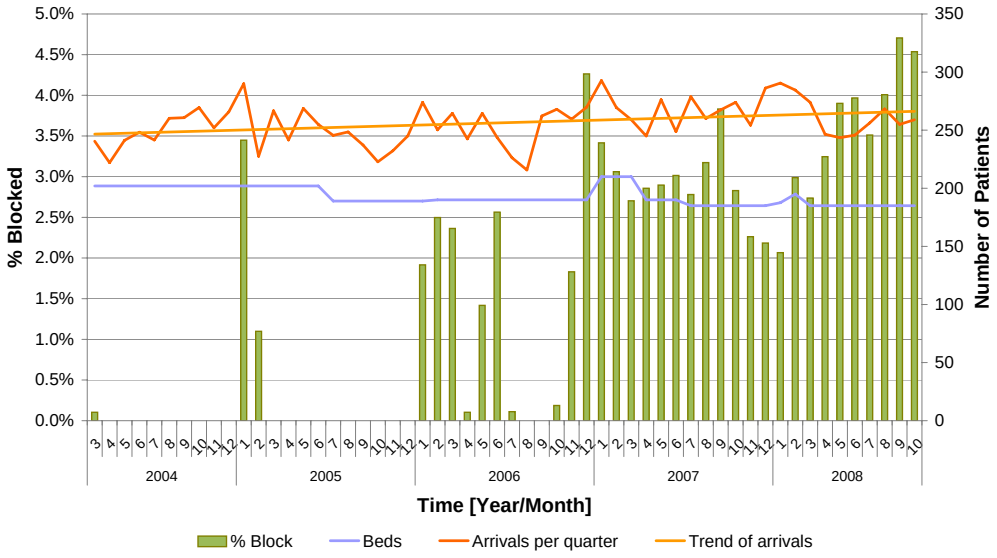


Figure 64: Percent of patients blocked, arrivals, and beds (2004-10/2008)

25 Departures from the Internal Wards

In this section, we describe departure rate patterns. These are different from the arrival rate patterns both in weekly and daily resolutions. Figure 65 shows weekly patterns. We observe that the departure rate during weekdays is higher than during weekends. There are very few departures on Saturdays. During weekdays, there are more departures on Sunday, Wednesday, and Thursday than on Monday and Tuesday. There is a tendency to release patients before the weekend, which is very clear in the data. This has two reasons: First, the hospital has fewer staff during weekends, who can therefore treat fewer patients; Second, there is the need to release beds for patients who are admitted at the beginning of the week. Figures 66 and 67 describe patterns at a daily resolution. We find that patients leave between 12:00 to 21:00, most of them between 15:00 to 16:00. The departures' daily pattern is very different from that of arrivals. This is due to the release procedure. Specifically, during morning hours, doctors check all patients and decide which ones to release. Then, between 11:00 to 12:00, they write release letters that they pass on to the nurses. The nurses continue the release process by guiding the patient and family through the release bureaucracy, and providing instructions for further treatment. If needed, the nurse coordinates the patients' transfer to another location (e.g. an elderly citizens' home or a rehabilitation center). This process takes several hours, and thus most patients are released in the early afternoon. The difference between the arrival rate and departure rate patterns determines the changes in the number of patients over the day and the unique LOS distribution that will be described below.

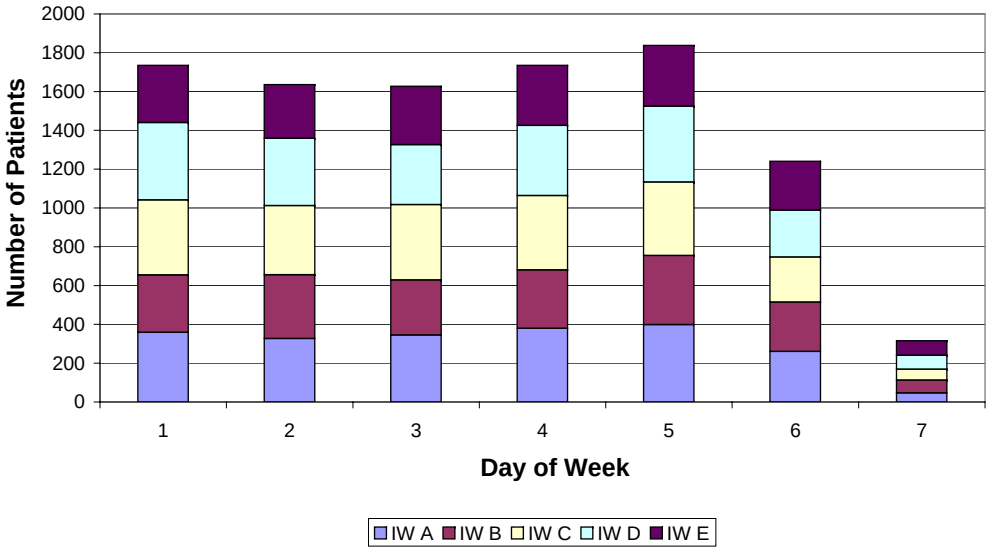


Figure 65: Departures by day in all Internal wards (2008)

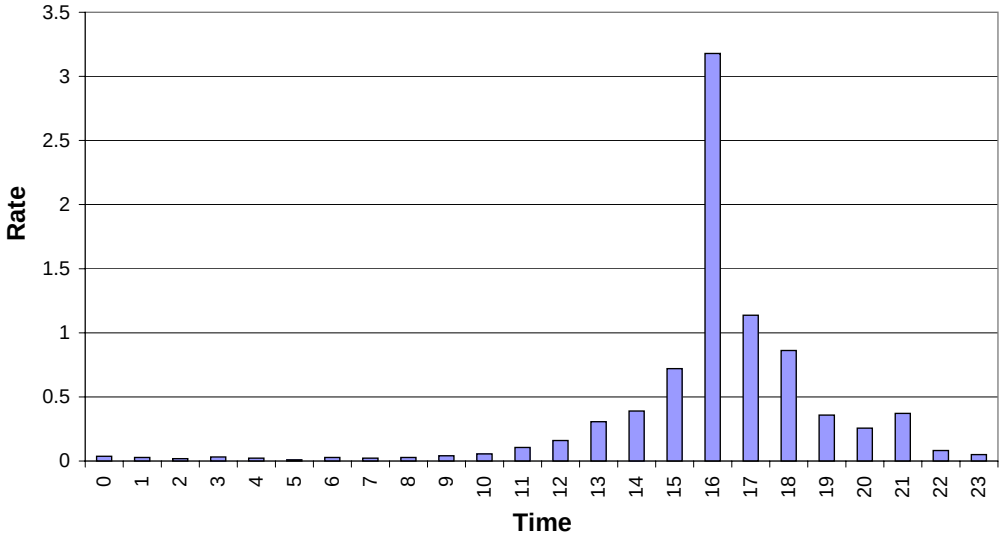


Figure 66: Departures by hour in weekdays at IW A (2008)

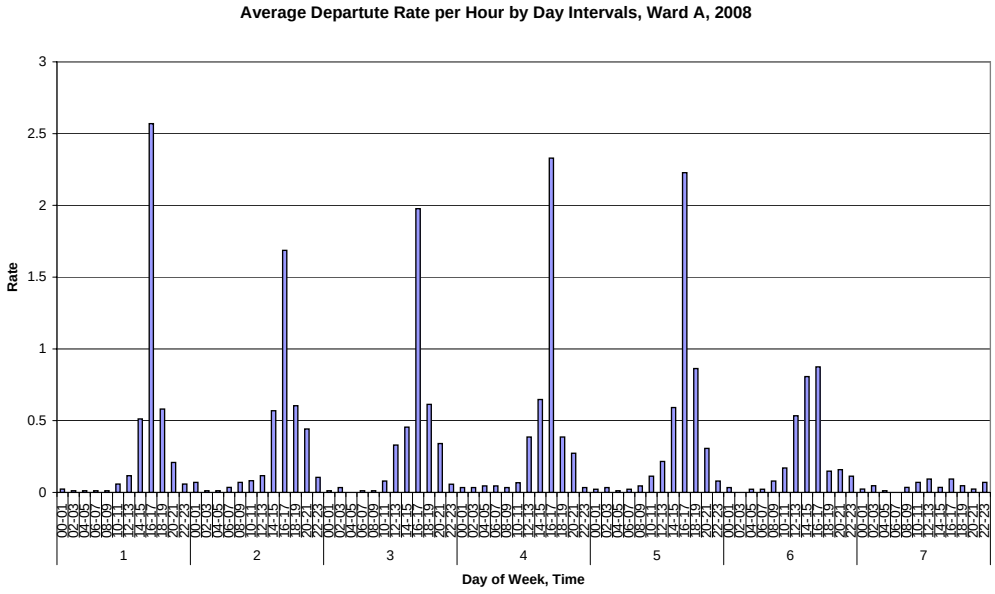


Figure 67: Departures by day and hour in IW A (2008)

25.1 Number of Patients (WIP) in the Internal Wards

In this section, we examine the behavior of in the number of patients (WIP) at different resolutions. Figure 68 illustrates the average number of patients in each IW, at a yearly resolution. We also indicated the number of beds on the same graph. This graph exhibits a very similar picture to the one we saw when examining the arrival rate. We observe the following interesting phenomena:

- During 7-8/2006, all the wards had a low number of patients. This was due to the war during that time. The low count is consistent with the decrease in arrival rates exhibited during this period, as seen in Figure 53.
- The number of patients in IW D decreased starting from April 2007. As we see in Figure 68, this is consistent with the reduction of the number of beds in this ward, starting at that time.
- Between 1-3/2007, the number of patients in IW B is very large. This is due to a unique organizational change that was implemented during that period, in which this ward was added a branch of 20 more beds, located in a different location.
- The number of patients in IW E was slightly reduced starting in 4/2008. This is again due to a reduction in the number of beds, which occurred because the ward was moved to a temporary location, with a smaller space, for renovation.
- In some cases, the number of patients is larger than the number of beds. In these situations patients are hospitalized in corridors.

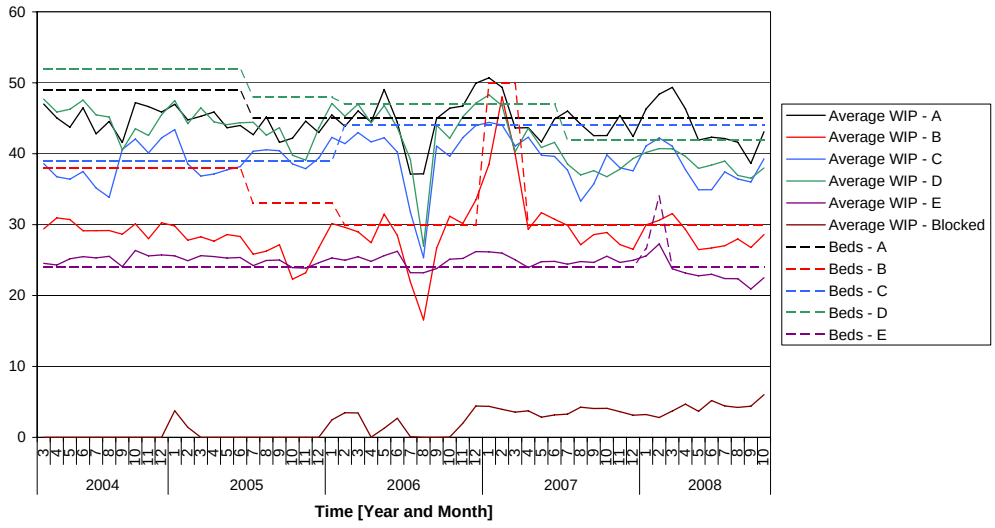


Figure 68: Average number of patients in each IW by month

When looking at a weekly resolution, for example in Figure 70, we note that the decrease in arrival rate and increase in departure rate, during Wednesdays and Thursdays, reduces dramatically the average number of patients towards the weekend. The fact that, during the weekend itself, patients are not released causes an increase in the number of patients that starts on Friday evening. When we study a daily resolution, we observe that the number of patients in the ward, during the afternoon hours, is significantly lower than the average number of patients during the whole day. Figure 69 illustrates this pattern in Ward A, by week days, and Figure 71 illustrates the same pattern in Ward A during the hours of the days on Sundays.

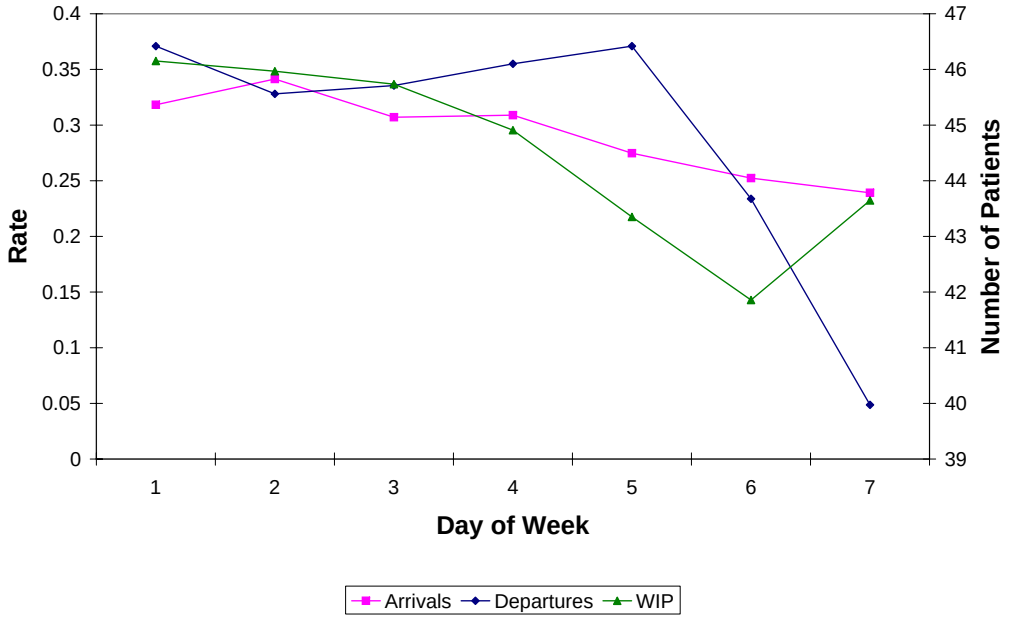


Figure 69: Arrivals, departures, and number of patients in Ward A by DOW

It is obvious that daily changes in WIP (patients count) imply daily changes in workload. As discussed in earlier parts of this thesis, it is customary to define personnel workload on the basis of number of patients in the ward. The fact that this is changing in time suggests that some adjustments in personnel levels should be performed during the day. The only problem with this idea is that workload is not evenly distributed over a patient's LOS. In reality, when patients are arriving to the ward and when leaving it, there is much more work for the nurse. If the arrival and departure rate were constant, this would not influence the workload. But when they do change overtime, they could have a dramatic impact on personnel workload. To understand this impact, we estimate the workload that each patient brings to the ward, using our Erlang-R model. Each patient alternates, during his stay, between "needy" and "content" states. When a patient is needy he requires service

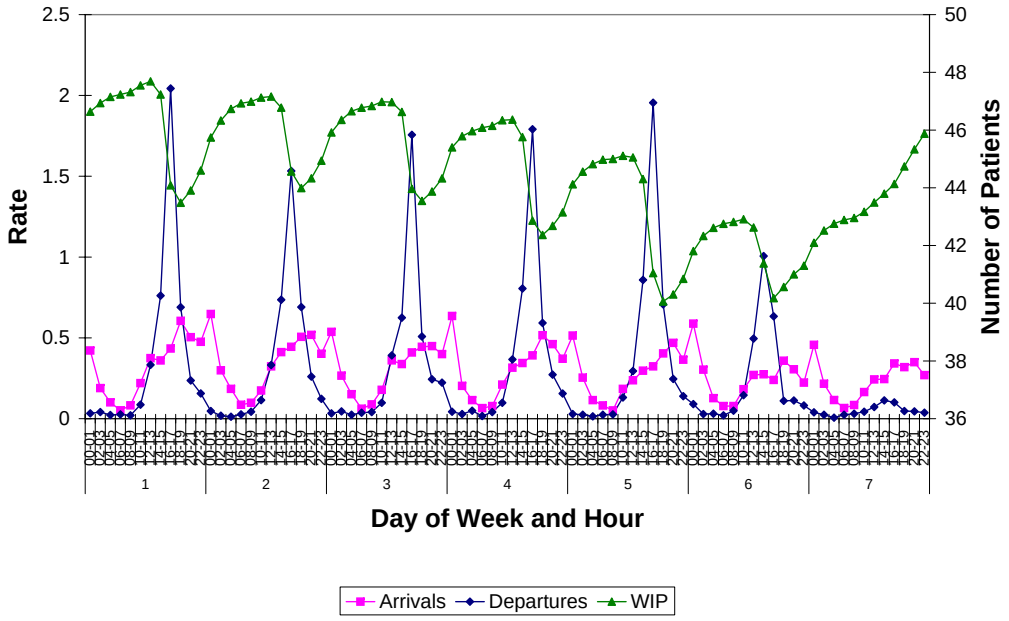


Figure 70: Arrivals, departures, and number of patients in Ward A by DOW and hour

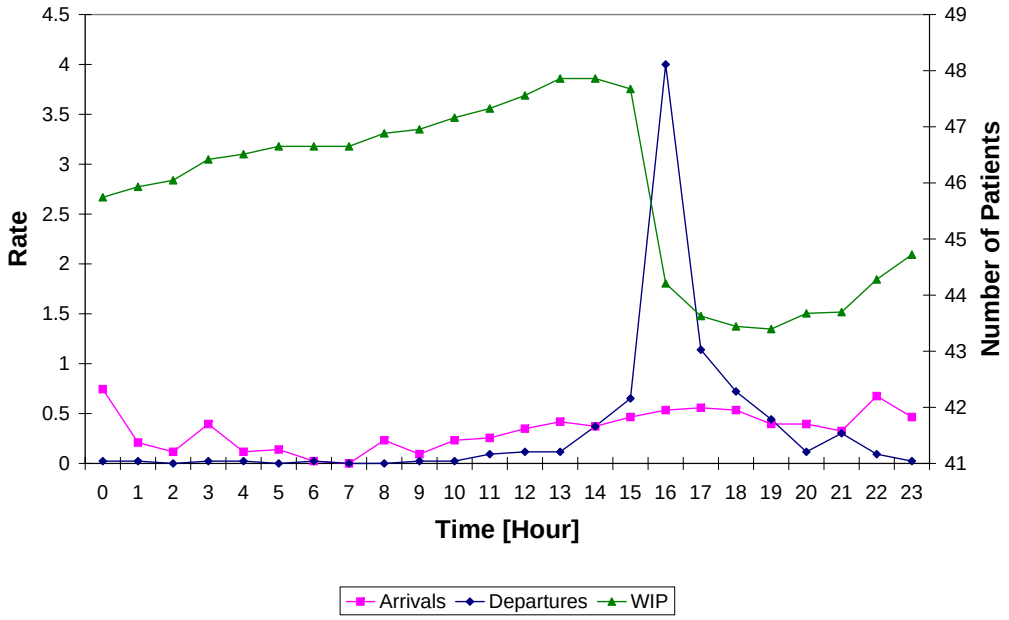


Figure 71: Arrivals, departures, and number of patients in Ward A by hour

from a nurse. A patient starts and ends his stay in a needy states. The average time of the first and last services (when admitted or discharged from the ward) could be different from the average service time of regular services (we do not have data on such times, but it is reasonable to make this assumption, and it is supported nurses). Thus, in order to define the average workload over the week, we shall count the number of patients that arrive, depart and are hospitalized during every hour, and multiply them by the service time they require. Figure 72 shows the changes in workload during one customer's stay. We see that the times when the WIP is lowest are those when the workload is highest. Thus, if one tries to calculate workload solely on the basis of the number of patients in the ward, one would significantly underestimate workload at the busiest time of the day.

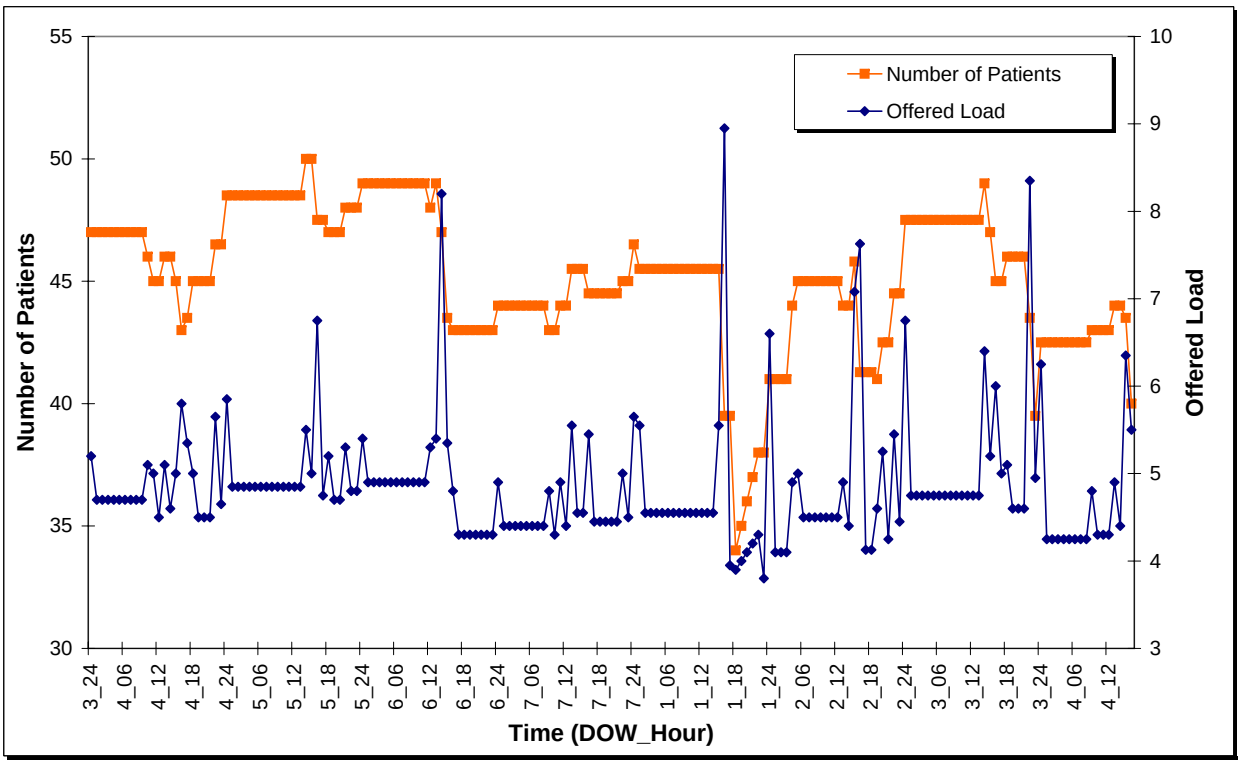


Figure 72: Number of patients and workload during the LOS of a random patient

One way to avoid this problem is to consider the number of patient at a specific time of day, and calculate personnel staffing at a weekly resolution. Figure 73 shows the distribution of the number of patients at midnight. We choose that time as the changes in the number of patients during night hours is low, and hospitals use it as a proxy for the load level in the medical ward. The distribution seems almost normal, which implies that classical queueing models might be useful for predicting the number of patients on a weekly resolution. One of the most important parameters for such

models is the LOS distribution, which we form to next.

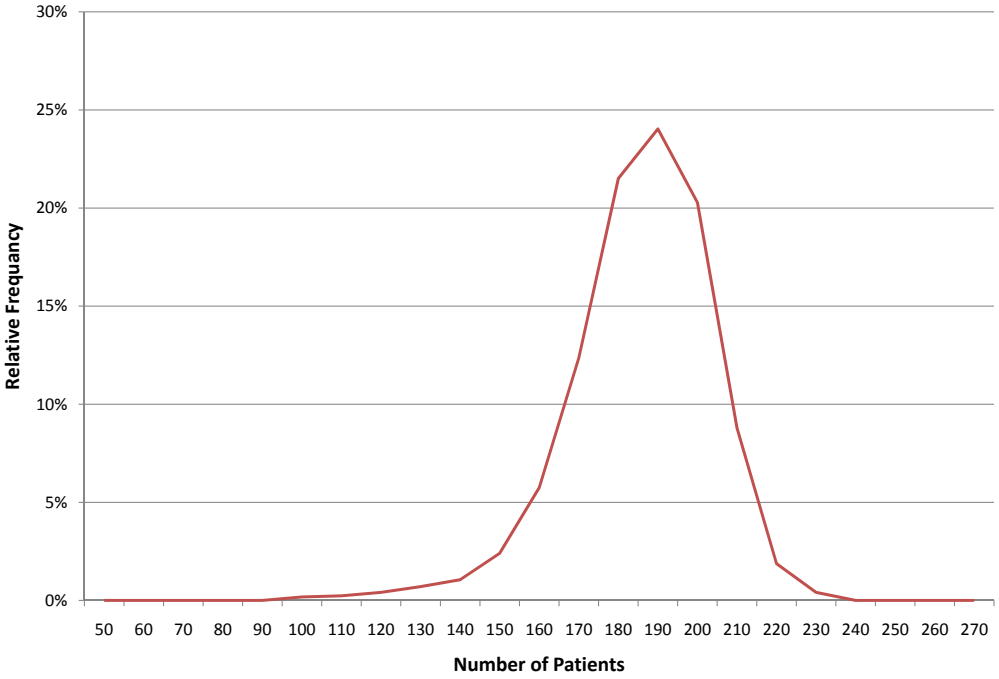


Figure 73: Density of the number of patients in all IWs

26 Length Of Stay (LOS) in the Internal Wards

In this section, we investigate the distribution of LOS, and the changes in ALOS during the years. As before, we first study yearly resolution to identify trends. Figures 74 and 75 depict the ALOS in all the IWs, at a resolution of years and month, respectively. We see that the ALOS has changed during the years and months, and is different among wards. It seems that IW E and B have much shorter ALOS. We found several explanations for that phenomena. IW E treats the easiest patients of all IWs, thus its patients are expected to stay a shorter time at the hospital. However, this is not the case for IW B. In order to understand this fact, we interviewed the management of IW B. In IW B, one of the doctors is assigned the task of reducing delays for exams and for specialist-services from other departments at the hospital. This is crucial at the beginning of a patient's stay, when the treatment plan is outlined. Sometimes ward B sends more patients per day to exams than is set by hospital regulations. These regulations were made to guarantee equal access to medical care by all wards, thus this practice might not work if all wards exercise it. When examining Figure 74 more closely, we also see that, during 1-2/2007, the ALOS of ward B was significantly increased. This was the time when IW B had the branch of 20 additional beds. This raises the conjecture that there is also an effect of the size of the ward, namely diseconomies-of-scale in our case: larger wards are harder to manage, which has a negative effect on the ALOS of patients. This conjecture is also supported by the reduction in the LOS in IW D from mid-2007, when ward size decreased. More research is needed, though it is clear that ALOS can be reduced. The fact that the LOS is not stable during the months and between wards raises the hypothesis that there are other parameters that influence ALOS. In the next section, we study one such hypothesis and investigate the connection between LOS and the workload in the ward.

We also examined the LOS distribution. Figure 76 shows the LOS cumulative distribution for the last two years. When looking at a resolution of days and hours we find two interesting patterns in the LOS distribution: We see that there is a clear stochastic order between wards A, C, D, and B. Figure 77 shows the LOS distribution in one of the Internal Wards. It can be seen that, when considering daily resolutions, the gamma and log-normal distributions fit the data well. The second and third graphs are at hourly resolution and illustrate the impact of the discharge policy: the decision on discharging patients is done once a day, hence the LOS distribution has peaks in spaces of 24 hours. This distribution looks like a mixture of several normal distributions. Figure 78 shows the LOS distribution for the other internal wards in a resolution of days.

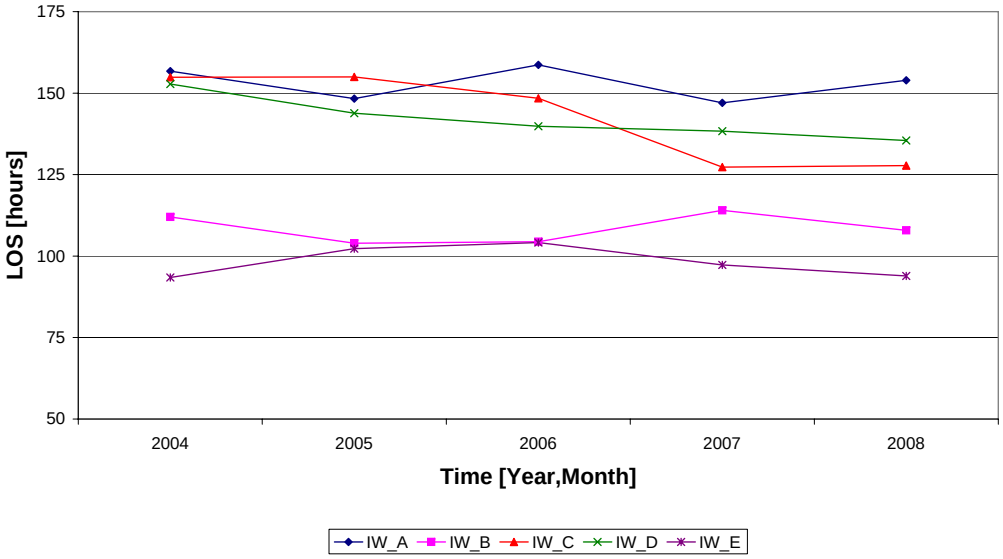


Figure 74: Average LOS in all Internal wards by year

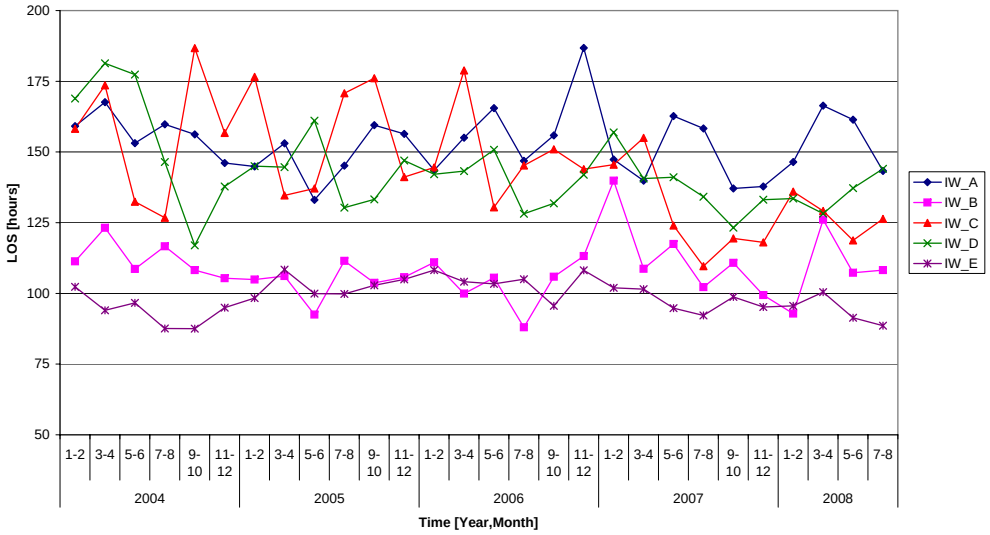


Figure 75: Average LOS in all Internal wards by month

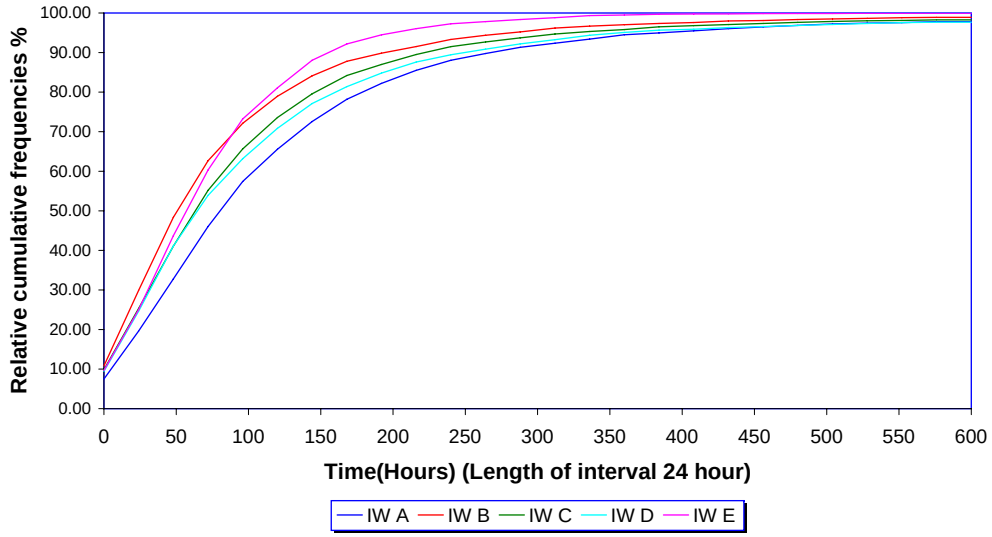


Figure 76: LOS cumulative distribution function of all Internal wards 2007-8

26.1 Length of Stay in Internal Wards as a Function of Workload

In this section, we investigate the impact of workload on the LOS in IWs. There could be two approaches to investigate this dependency. The first is using econometric tools as in [20], the second using queueing theory as in [56]. We follow the latter as it provides a more complete picture of the relation between the two measures. There are several ways to define workload. We consider the patients load, i.e. load due to the number of patients present in the ward. If LOS depends on the workload, we expect to find that, in highly loaded times, the release rate from the ward is higher then in lightly loaded times. This higher rate could be a positive effect that represents the ability of the system to increase its efficiency when needed, or a negative sign of overcrowding that enforces doctors to send patients home too early. One can distinguish between the two, for example by examining the fraction of returns as a function of workload.

In order to find whether release rate depends on the number of patients in the ward (denoted by l), we consider the medical ward as a black-box and fit a birth and death process to its number of patients. We then examine the fit of several queueing models to the data. The models we consider are M/M/∞, M/M/s, and M/M/s/n. In these models, it is implicitly assumed that there is no dependency between the release rate and the state of the system. Hence, in all of them there is an increasing linear relation between the death rate and the number of customers in the system (at least over some range of l). Figure 79 demonstrates this relation.

In order to fit a birth and death process to the number of patients in a ward, we use the following notation: Define T_l to be the average time that the system is in state l (the ward has l patients)

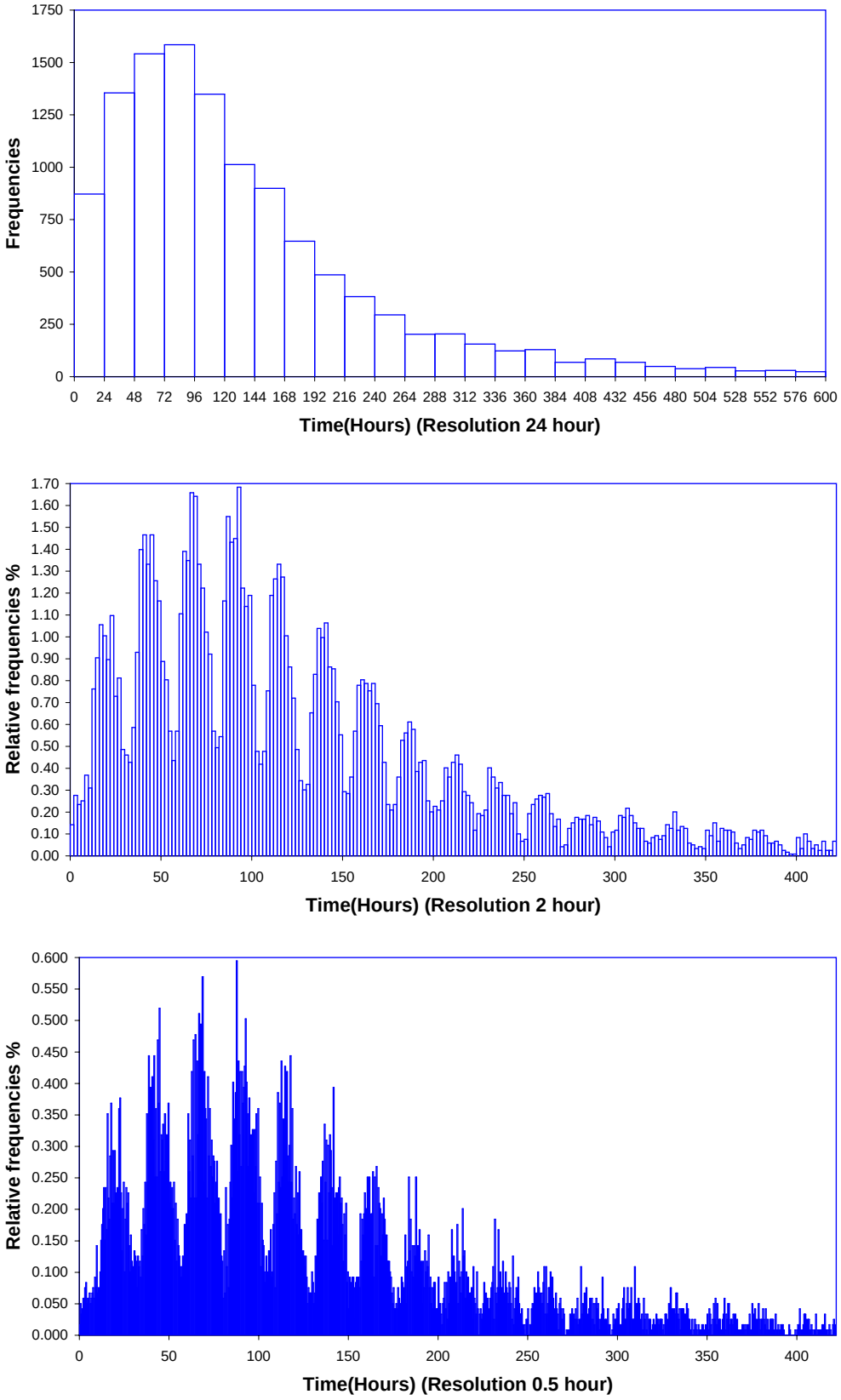


Figure 77: LOS distribution of IW A in several time scales

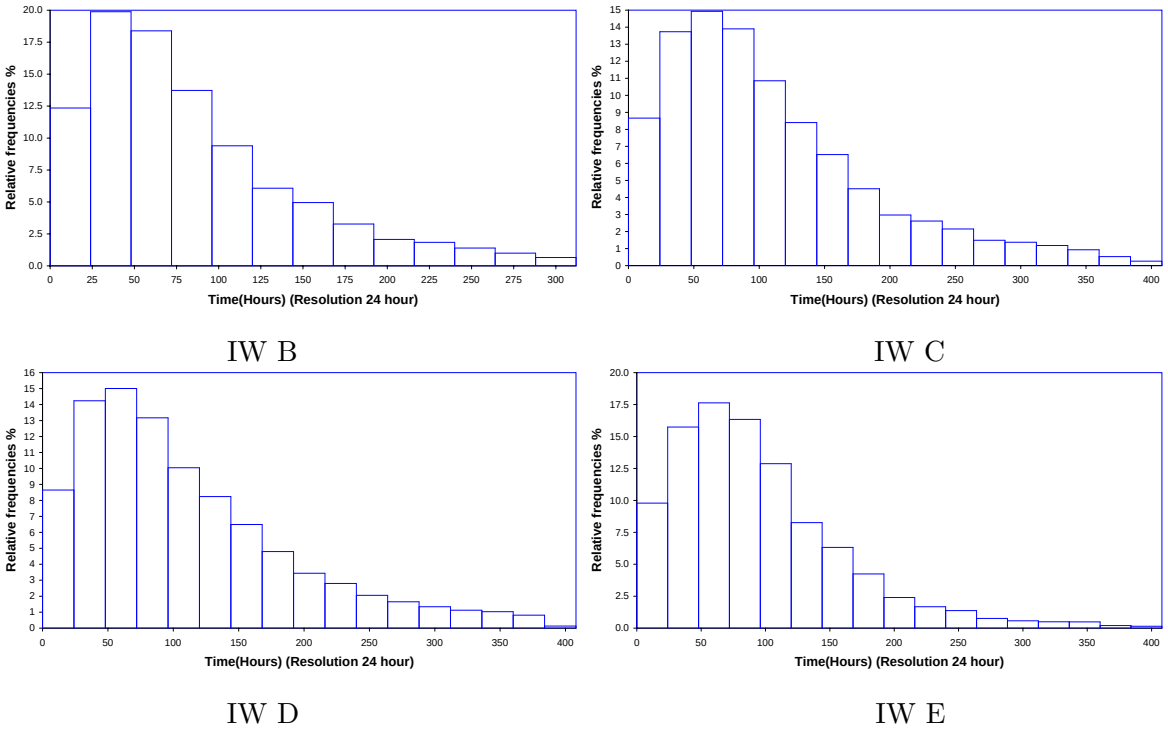


Figure 78: LOS distribution of IW B-E

until it moves to either state $l+1$ or $l-1$ (i.e. a patient enters or exits the ward). In addition, define $P_{l,l+1}$ as the probability to transfer from state l to state $l+1$, given that it was in state l, and $P_{l,l-1}$ as the probability to transfer from state l to state $l-1$, given that it was in state l. We estimate these probabilities by calculating the proportion of times the system moved between these states, given that it was in state l. Then, $\frac{P_{l,l+1}}{T_l}$ and $\frac{P_{l,l-1}}{T_l}$ are the birth and death rates from state l, respectively. We then examine the death rate as a function of l, in order to support or refute our hypothesis, i.e. does one of the classical queueing models fit the data?.

Figure 80 depict the death rate as a function of l, for all wards during the period 8/2007-7/2008. We chose that period since there were no changes in the size of wards, and we are unaware of any policy changes during that period. The range of states varies among wards because each Internal ward has different bed capacity. We note that death rate is constant for most system states. In some of the wards, (e.g. IW B and E) the death rate seems do decrease in the first lowest states, and in other wards (e.g. IW C and E) we note an increase of the death rate in the highest states. As can be seen in Figure 81, this almost constant death rate means that the release rate from the ward decreases as the number of patients in the ward increases, which is different from previously mentioned queueing models. The increase at the end means that if the ward is overcrowded than patients are actually sent home, in order to clear beds in those wards.

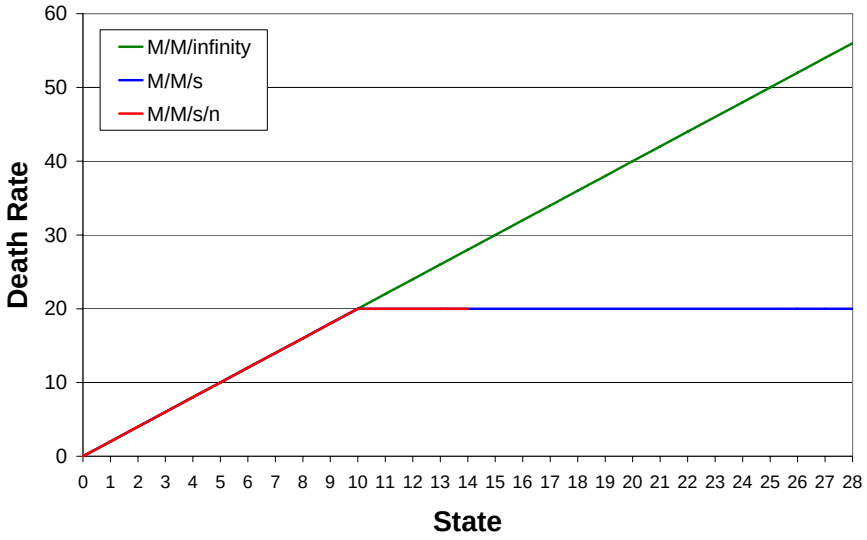


Figure 79: Theoretical Relation between death rate and system's state

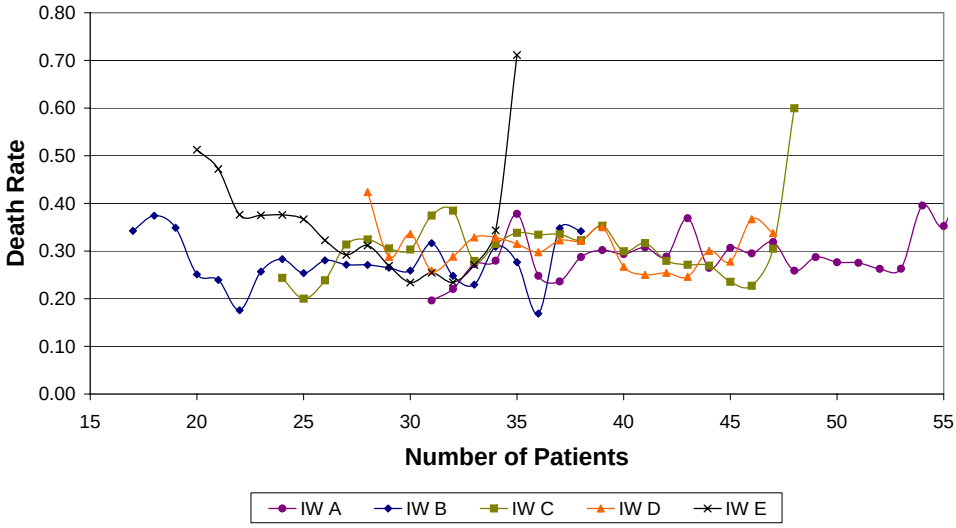


Figure 80: Death rate as a function of number of patients in ward

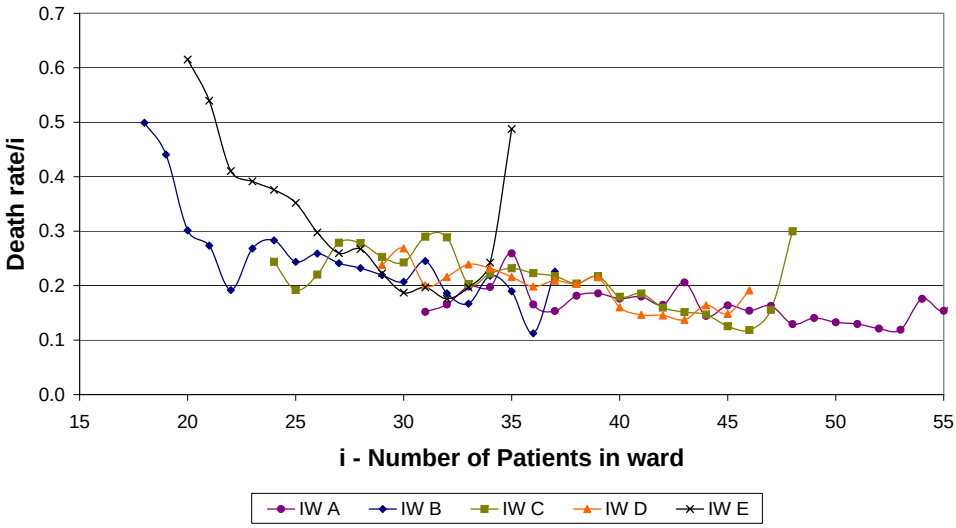


Figure 81: Release rate as a function of number of patients in ward

26.2 Simulating Internal Wards Length of Stay

In Section 26, we described the LOS distribution in two time scale: days and hours. The latter distinct shape can be created in the following ways:

- Method 1: Create hourly LOS as a mixture of normal distributions.
- Method 2: Sample the number of days from Log-normal distribution and the actual departure time from a normal distribution around 15:00.

Using the second method is not simple: When creating an IW simulation, this two-steps procedure of sampling LOS must be combined with the dynamics of the IW itself, and thus one would like to create a mechanism in the simulation in a way that eventually results in this kind of LOS.

We propose the following mechanism: After arrival, sample a departure time for the following day around 15:00. This time is normally distributed. Then determine departure by drawing from the following distribution: given a patient is on her t-day of hospitalization, release the patient with probability h_t and stay for another day with probability of $1-h_t$; Here the values of h_t are drawn from the hazard-rate function of the LOS distribution in a time scale of days ($h_t = P(LOS = t|LOS \geq t)$. Figure 82 shows this hazard-rate function for IW A. It seems that h_t is increasing at the beginning, then decreasing, and finally stabilizing. Figure 83 shows the LOS created by this method verses actual LOS data in IW A.

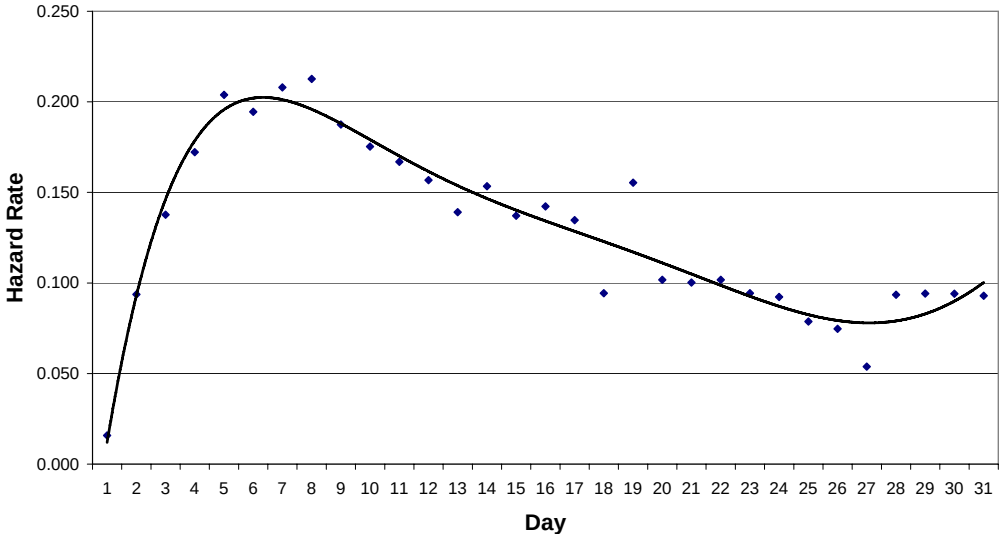


Figure 82: Hazard rate of LOS in IW A

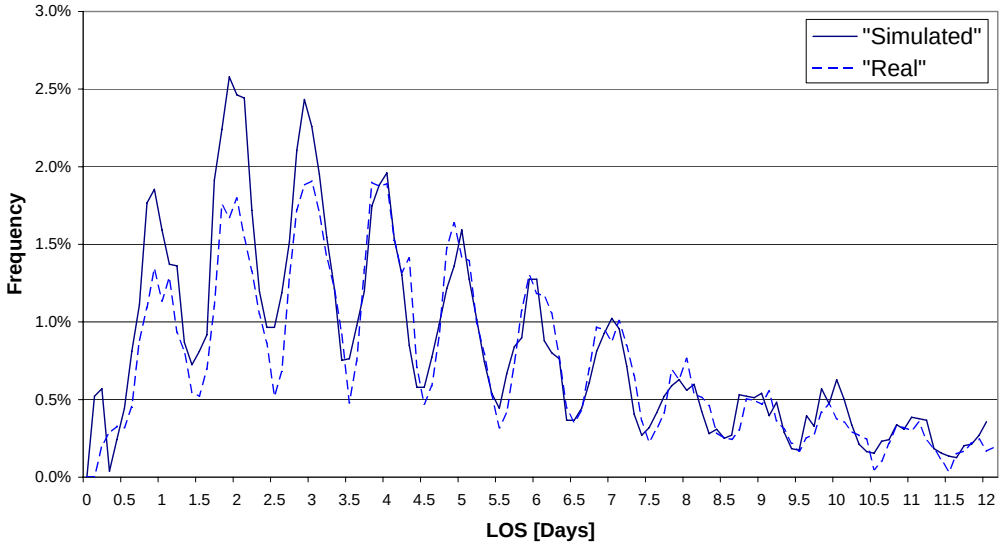


Figure 83: Simulated LOS vs. real data

27 Returning to Hospitalization - Internal Wards vs. Oncology Wards

Up to now, we concentrated on IWs, but a hospital is rich with other medical wards that have different characteristics. We now study the phenomenon of returning to hospitalization, as a classifying characteristic among wards, and discuss its significance for determining beds allocation. We distinguish between two types of wards: First are wards where a patient is admitted for a one-time treatment, such as an IW. In such wards, if a patient returns shortly after a previous visit, the return is considered negatively, possibly due to a lack of treatment received (though this is certainly not always the case). The second type of wards, such as an Oncology ward, is one in which patients return for treatment every several weeks. In this case, patients are expected to return, and each patient visits the hospital many times until cured. Table 5 compares the average number of returns per patient, in the studied period, and the probability of return within three months. We see that the two types of wards are indeed very different.

Ward	Average returns per patient	Average time between successive returns of a patient (days)	Probability of return within 3 month	ALOS (days)
Internal	1.76	208	22%	4.8
Oncology	5.46	22	75%	3.4

Table 5: Returns to hospital

Figure 84 shows a histogram of the number of visits per patient in Oncology wards. Figure 85 shows the distribution of times between successive returns of patients in the Oncology wards.

This analysis improves our understanding of the connection between the number of beds needed and the corresponding offered load: the latter requires a separation between hospitalized patients and those who are currently on "vacation" (i.e. at home). While in an IW, one can generally assume that each visit of a patient is independent of previous visits, this need not be the case in wards such as Oncology: the latter must reserve space for patients that are in the middle of a series of treatments. In fact, these patients must have a higher priority over new patients, who can be transfered to another facility if needed. Hence, Oncology planning is closer to the planning of a medical clinic (see Green et al. [37]).

We propose to use the Erlang-R model to solve the bed allocation problem of such wards. In this case, "Needy" patients are those who are currently hospitalized in the ward, "Content" patients are the ones who are under the responsibility of the clinic but currently at home. The model enables

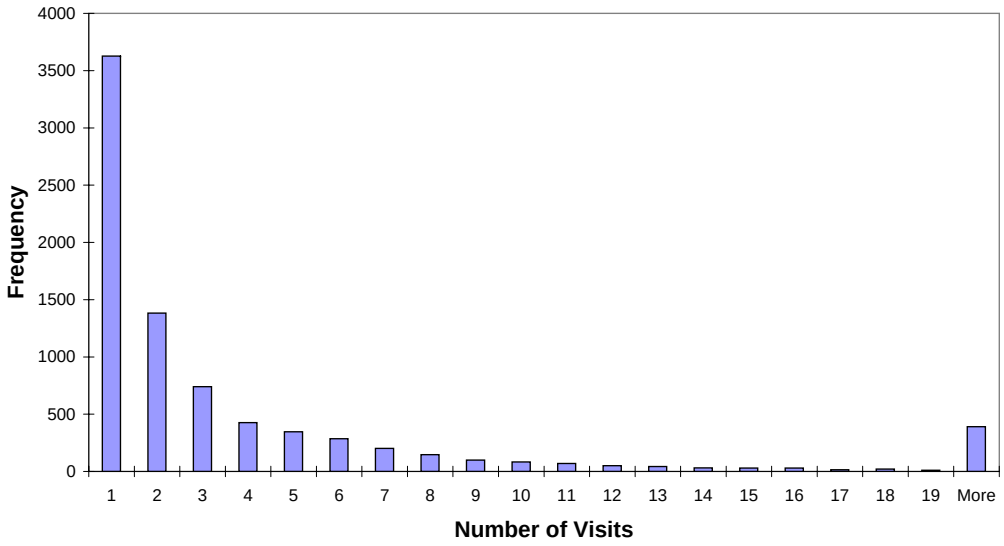


Figure 84: Number of visits per patient in Oncology wards

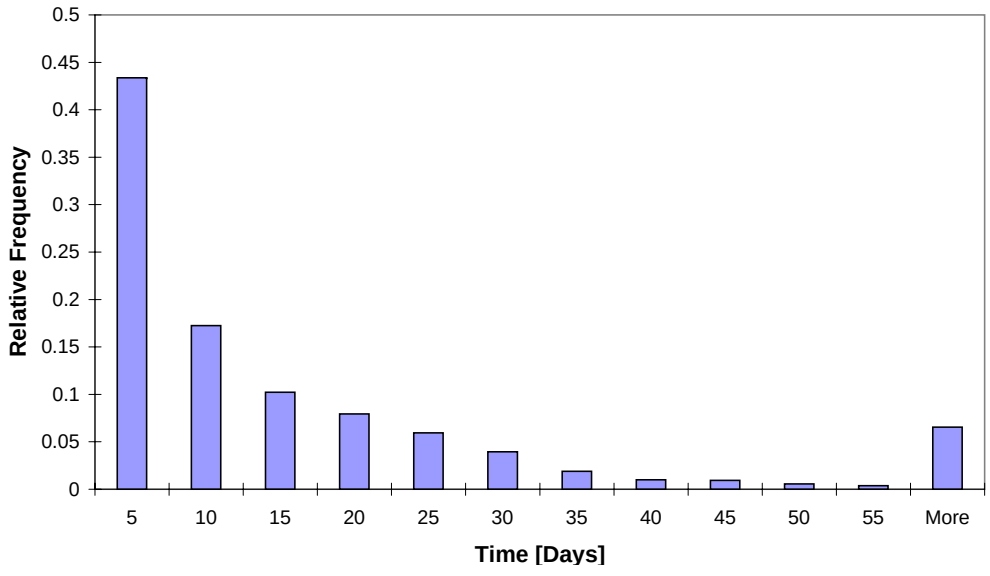


Figure 85: Distribution of times between successive visits of patients in Oncology wards (5 returns or more)

one to separate the two streams of customers: new and returning, and set each one its own service goals. Here s will be the number of beds allocated to the ward, while n can set the maximum number of patients treated by the ward. Returning patients should not wait long for a bed. On the other hand, treatment is very expensive, which suggests that in this case $P(W > 0)$ should be lower then P(block).

The offered load ratio, in Oncology wards, is $B = \frac{1/22}{0.817 \cdot 1/3.42} = 0.19$. Using our analysis in Section 19.1.3, we deduce that with such offered load ratios, considering the influence of returning customers is very important. Therefore, the semi-open Erlang-R model is more suitable than the Erlang-B model (as suggested by de Bruin et al. [10]). There is also another difference between the approach of de Bruin et al. and ours, as they have not made the distinction between the number of beds (s) and the maximum number of treated patients (n), and assumed them to be equal (i.e. $s = n$).

Figure 86 shows the arrival rate per month to Oncology wards during 2006-2008. Note that there is an increasing trend in the arrival process. We observe an increase of 11% per year in the number of visits to the Oncology wards. This fact suggests that it might even needed to use a time-varying model.

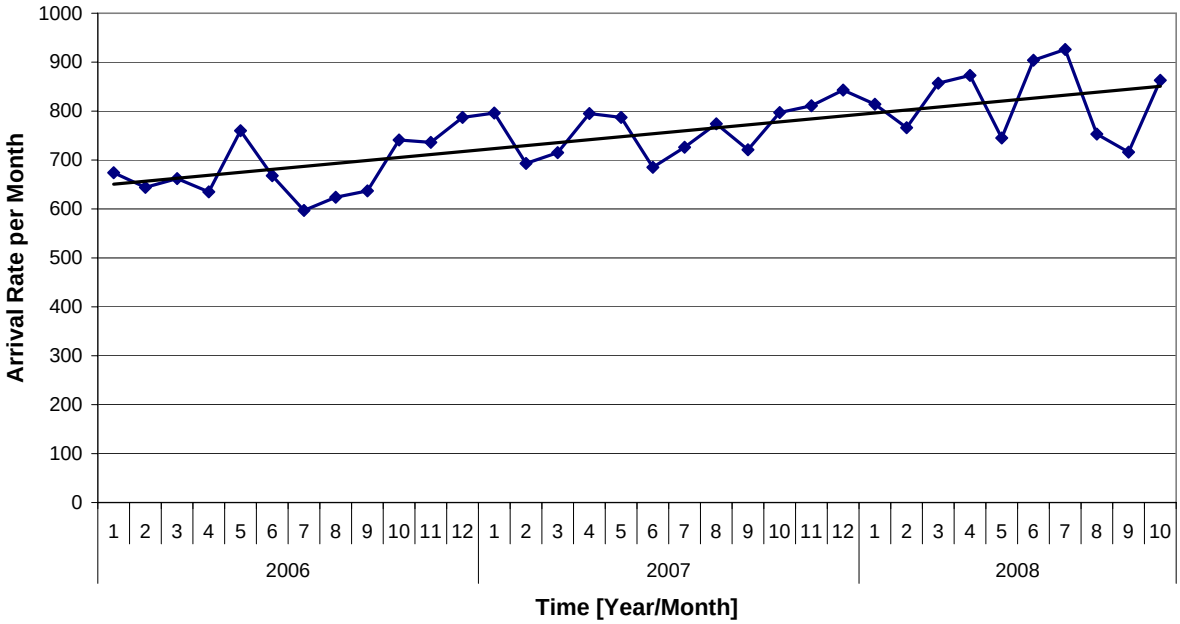


Figure 86: Arrival rate to Oncology wards, 2006-2008

28 In What Regime Do the Internal Wards Operate?

In the field of service engineering, it has become customary to distinguish between three operational regimes: The efficiency-driven regime (ED-regime), the Quality-driven regime (QD-regime), and the Quality- and efficiency-driven regime (QED-regime). The ED-regime emphasizes the efficiency of the system: servers are highly utilized (close to 100%), and hence customers typically suffer through a long wait for service. In the QD-regime, the emphasis is on the quality of service provided to served customers: servers are available for service for a significant part of the time, and customers hardly wait for service. The QED-regime is somewhere in between: the emphasis is both on service quality and on servers' utilization. In large systems that operate in this regime, we find that servers' utilization is around 90% and customers wait about half of the times. We would like to determine in which of these regimes the IWs operates. We discuss this in relation to beds and doctors.

We argue that beds' capacity of the IWs is managed in the QED-regime. To support this statement, and in view of the fact that we have no data on the nurses operation (service time and staffing levels), we propose to fit a loss model (Erlang-B, as in de Bruin et al. [18]) to calculate a theoretical value for the probability of blocking. The probability of blocking is the probability that a patient does not find an available bed in the unit he needs.

In the QED-regime ($n \approx R + \beta\sqrt{R}$), $P(block) \approx \frac{\gamma}{\sqrt{n}} = \frac{1}{\sqrt{n}} \frac{\phi(\beta)}{\Phi(\beta)}$, and the occupancy level $\rho \approx 1 - \frac{\beta+\gamma}{\sqrt{n}}$, both are $O(1/\sqrt{n})$. Taking the data of year 2008, we find that the average LOS in all IWs is 5.12 days, there were 186 beds in all the IWs, and the total arrival rate is 34.4 patients per day. Thus $\beta = \frac{n-R}{\sqrt{R}} = \frac{186-34.4*5.12}{\sqrt{34.4*5.12}} = 0.4$ and, $P(block) \approx 2.9\%$. We checked, in our data, the fraction of patients that were physically hospitalized in other wards, but were still under the medical responsibility of IW doctors. We found that it is 3.54% of the patients, which is quite close to the the theoretical value of 2.9%. In addition, the approximation for the occupancy level is $\rho \approx 91.7\%$, which is again very close to the actual value of 93.1%. Therefore, the facts that support our conclusion are:

1. The blocking probability is 3.54%.

2. Average beds' utilization is 93.1%.

Note that fitting a semi-open Erlang-R (assuming one had the data on nurses) would have retained the above conclusion, due to the orders of magnitude of blocking and utilization: both are of QED order.

When considering doctors, we argue that doctors operate in the ED-regime. This claim is based

on the following procedure: from 16:00 in the afternoon to 8:00 in the following morning, there is only one doctor on duty in each IWs. This doctor admits most of the new patients of the day. Therefore, if a patient is admitted to an IW (i.e., only if there is an available bed) he must wait until both a nurse and the doctor on call are available. The average reception time by the doctors is 30 minutes. Thus, the appropriate model for considering the waiting of a patient in the Emergency ward until transfered to one of the IWs is an $M_t/G/1$ model. We hypothesize that doctors operate in the ED-regime, since the service time is 30 minutes, while the waiting time of a patient in the Emergency ward has an average of 2.5 hours (see Tseytlin [71]). This is a common characteristic of the ED-regime, where the waiting is much longer than service time.

29 Part III - Conclusion and Future Research

In this part of the thesis, we have observed interesting phenomena obtained through extensive empirical analysis of patients flow data. We showed how arrival and departure rates change in time, and created distinct LOS distribution. We showed how arrivals and departures combine to create changes in the number of patients, and the implication of that observation on staffing and bed allocation. We discussed several interesting problems that arise from the data, such as the dependency between load and LOS.

Our data is very rich, and much more analysis is called for. As an example, we plan such empirical analysis for two Maternity wards, operated in parallel, where one distinguishes between three types of patients: normal birth, pre-birth (pregnancy) complications, and during-birth complications (typically due to a Caesarean section). We would like to investigate whether this distinction has an impact on LOS distribution and how it effects load balancing problems between the two wards.

Part IV

Future Research

In this last part, we propose several possibilities for future research.

30 Combining Managerial / Psychological / Informational Diseconomies of Scale Effects

In their article Boudreau et al. [12] discussed the importance of combining Operation Management and Human Resource Management (HRM) models. The integration between the two fields is very challenging. This integration requires some collaboration with researchers from the field of HRM, which is not common practice. We consider it important to try and carry out such integration in nurse staffing for one main reason; when discussing our models with doctors from an Israeli Medial Center, there was major concern that the model might recommend "too large" departments. The claim was that there is a managerial / psychological / informational limit to the number of patients one MU can actually treat, and if that limit is surpassed, the quality-of-care deteriorates. The claim was that a large MU is inferior to several small MUs, even though in a smaller MU one has fewer medical personnel. We found that claim interesting, and suggest a few explanations for the source of this diseconomies-of-scale effect.

1. *Managerial causes.* The MU manager is the one most responsible for the medical decisions. The doctors work as a team, consulting one another about the patients before taking a medical decision. Adding more doctors thus increases the level of knowledge of the MU but, due to the form of responsibility distribution, it might not always lead to a similar rise in capacity. The framework of team vs. individual responsibility has been investigated in the field of HRM. In our context, one can find the two ways of setting, i.e. (a) team vs. (b) individual. (a) A team of doctors was mentioned above. One can say that nurses also work as a team if all of them are jointly responsible for all the patients in the ward. (b) In the individual setting, there is one nurse for one or more specific patient(s) as in the case of Intensive Care Units. From the operational point of view, the two settings are different.

2. *Psychological causes.* Are there unique learning and forgetting effects? It seems that there is a difference in learning schemes between various service environments, such as the machine-repairman problem, call center and nursing staffing problems. We know that in call centers,

each customer is different, though his/her problems are alike (and can be classified into certain types). One can see a learning curve in which as a new service person starts to work she begins to learn the different types of services; as she deals with more customers she becomes more professional which, in turn, is reflected in the rise in the quality and the efficiency of the given services. This learning effect is known also in Industrial Engineering. This means that, in reality, the service rate μ is not a constant parameter but is actually a function of time, i.e. $\mu(t)$. Usually we ignore this effect by looking at steady state, assuming that all nurses have sufficient experience. In repair-man and nurse-staffing problems we see the same effect: as one service person deals with more customers (i.e. machines or patients) she learns more, becoming an expert. An expert will deal with problems more efficiently; she is capable of treating more customers better and faster. But there is one difference: in nursing and machine-repair problems each customer "calls" the server several times during his stay/use, but the server needs to treat each customer as an individual and to remember his problems. This is customer learning effect. Thus, as the number of patients per nurse increases we might find a dis-economic-of-scale effect, where the reset-time of starting the treatment of a patient might grow with the number of customers. One could say that μ is actually a function of the system, for example: $\mu(s, n)$. One might be able to make some assumptions on the shape of that function (e.g. convex or constant). Combining the two learning effects, the customer learning and the job learning, one may define a service function $\mu(s, n, t)$. But as mentioned before, we may prefer to look at steady state, and assume that all nurses have sufficient experience, and ignore the time effect. The shape of the function $\mu(s, n)$ is not clear; it might increase due to overload effects (see below) or decrease as usually happens in learning curves.

3. *Informational causes.* How does the medical personnel react to the information overload caused by a large number of patient? Hall and Walton [40] reviewed some literature on Information overload in Health-Care systems, which raises some possible effects of overload. This raises the following question: Does the number of errors rise as a function of the nurse-to-patient ratio or as a function of the unit size?

These possible explanations could be investigated; with the combination of the Mental- and Physical-capacity into a single model is being important and challenging. Technically speaking, if we could define some Quality-of-Service (QoS) function in one of the following ways: (a) As a function of the workload (b) As a function of the number of patients in the system, i.e. $f(n)$, or (c) As a function of the number of servers and patients in the system, i.e. $f(s, n)$, then we could

combine this QoS function into our optimization model. The following ways are possible:

Define $E(QoS)$ - the average Quality of Service in the MU with n beds and s nurses, or define $P(QoS < \alpha)$ - the probability that the service quality in the MU will be less than α. Then,

$$
\begin{aligned}
& min_{n,s}\ C(n,s) = C_n n + C_s s; \\
& s.t\ \ P(W > t) < a; \\
& \quad P(block) < b; \\
& \quad P(QoS < \alpha) < c;\ \text{or}\ E(QoS) < c \\
& 0 \le s \le n;
\end{aligned}
$$

The calculation of the QoS measure will be based on the system product-form solution, or its approximations.

31 Phases of Treatment or Heterogeneous Patients

31.1 Combining the Phases of Treatment During the Hospitalization Period

When discussing this research with some medical crew, we became aware of an interesting phenomenon; at the beginning of the hospitalization period (approximately the first twenty-four hours), the patient requires intensive care while, as days pass, the care becomes less and less intensive. One can model this fact as a frequency reduction or as a service-duration reduction over time (i.e. the service-time function will decrease as a function of the LOS). We can divide the stay into a finite number of sequences, and categorize them. I will demonstrate this with the two following classes: (A) Intensive Care and (B) Regular Care. The following Figures, 87 and 88, illustrate two possible models of the system. The first assumes that a patient can move from class A to B and the reverse. The second model assumes that a patient starts in class A and, at some point in time, he moves to class B and later leaves the system from class B. In both models the differentiation between the stages was modeled through frequency reduction. This was achieved by using the following assumption: $\delta > \gamma$.

The first system is a Jackson network, with the same structure as described in Chapter 17. Thus, it can be solved and approximated using the same methods. (open questions: is there an assumption on the relationship between the sorting probabilities? how can we calculate them?) The second suggestion is a closed BCMP network[4] (which has a product-form solution defined by

[4]BCMP network contains an arbitrary but finite number N of service centers, and an arbitrary but finite number R of different classes of customers. Customers travel through the network and change class according to transition

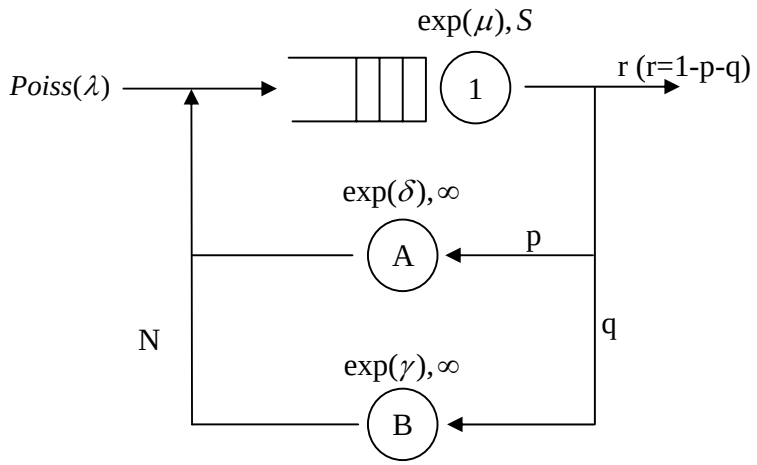


Figure 87: Phases of hospitalization - Model 1

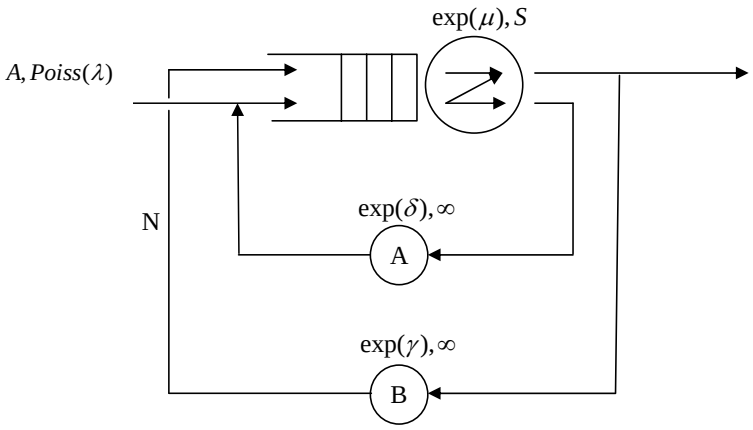


Figure 88: Phases of hospitalization - Model 2

Baskett et al. [9]), assuming FCFS discipline with identical and exponential service rates. This model describes the situation more precisely, but it might need to be solved separately.

Note: This is one of the differences between the call center or repairman problem and nurse staffing. Is it similar to a learning effect in some sense? (the learning of specific customer instead of the learning of the server)

31.2 The Influence of Time Delays Before and After Medical Analysis or Surgery

In addition to the description of classes of patients in the previous subsection, one can add classes regarding the treatment stages, i.e. before and after medical analysis or surgery. The model shown in Figure 88 also fits this situation. This modeling allows us to give priority in medical care to each class of patients, by setting different waiting threshold for each class. The questions that arise here are about priority schemes, and staffing levels.

One might also want to look at a more delicate issue, and check with doctors the possible influence of time delays on the patient situation. Is there a change in the rate of patient's state during the waiting time, as a function of the patient's class? Our assumption throughout this work was that this rate is constant, i.e. there is a linear connection between the waiting and the patient's state. But if it is not, for example, if it is an increasing function, then the model might give an underestimation for the required staffing levels. (Remark: in the case of the machine-repairman problem it could be either way, depending on the mechanical aspects of the machine; there are "delicate" machines that need repairing immediately, while other machines can stay for weeks without changing their state). This function could also change over stages, which means that when a patient's state is critical, the rate might be increasing, but when the medical state of the patient is stabilized (i.e. not critical) the rate might be constant. (For a model that combines the changing influences of waiting and service (in the psychological aspects) see Carmon et al. [14]).

probabilities. Thus, a customer of class r who completes service at service center i will next require service at center j in class s with a certain probability denoted $P_{i,r,j,s}$. There are four types of BCMP networks, that differ according to several condition it satisfies concerning the service discipline and service-time distribution. We will consider here only Type 1. The conditions are: The service discipline is FCFS; all customers have the same service time distribution at this service center, and the service time distribution is a negative exponential. The service rate can be state-dependent where $\mu(j)$ will denote the service rate with j customers at the center.

31.3 Classes of Patients (Heterogeneous Patients)

There are situations where the MU treats a few types of patient, who could be divided into several groups (or classes) that are independent of each other. There are situations where the classes of patients are dependent in various ways; we dealt with such models in the sections above. When the classes are independent, the model is simpler. An example of a two-class model, can be viewed in Figure 89. The network is a simple BCMP network, where patients from one class cannot transfer to another class but stay in the same class for their complete stay in the system.

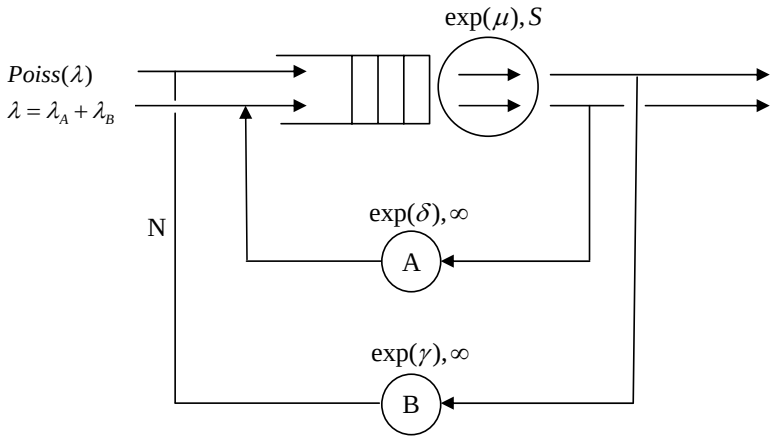


Figure 89: New model for two classes of patients

32 Nurses in the QED Regime, and Doctors in the ED Regime

There is a difference between the cost of doctors and of nurses. This difference suggests that a different regime must be considered when setting staffing levels for each type. The more expensive resource should be ED-staffed, gaining very high utilization levels, while the less expensive resource should be QED-staffed, with a balance between utilization and service level. We can model this situation in the following way: we assume that nurses and doctors are separately required, therefore, the communication and synchronization between nurses and doctors are made at separate times, i.e. not in front of the patient. The model contains two stations of the M/M/S type with s_N and s_D servers (denote as node N for Nurses, and node D for Doctors), and one M/M/∞ node, as seen in Figure 90. In this way it is a closed network, with a fixed population of n patients. The situation could have been modeled also as a semi-open network that contains one entrance node of the type M/M/1, as done with the model we analyzed previously (in Chapter 11). In this case we believe,

based on our previous experience and knowledge of the ED and QED regimes, that the appropriate QED+ED conditions might be:

$$\lim_{\lambda\to\infty}\frac{n-s_N-s_D-\rho_1}{\sqrt{\rho_1}}=\eta, \qquad -\infty<\eta<\infty; \tag{i}$$

$$\lim_{\lambda\to\infty}\sqrt{s_N}\left(1-\frac{R_N}{s_1}\right)=\beta_1, \qquad -\infty<\beta_1<\infty; \tag{ii}$$

$$\lim_{\lambda\to\infty}s_D\left(1-\frac{R_D}{s_2}\right)=\beta_2, \qquad -\infty<\beta_2<\infty. \tag{iii}$$

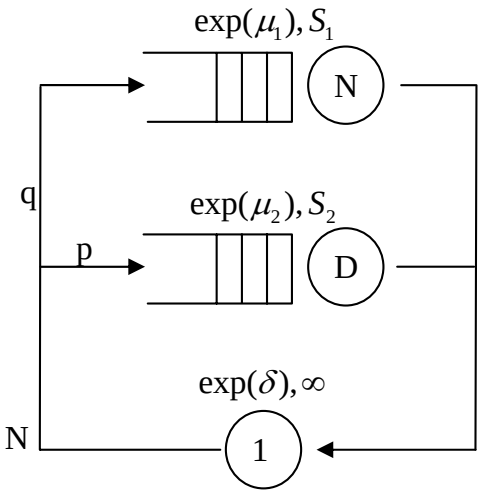


Figure 90: ED (doctors) and QED (nurses) model

Another option is to assume that, in some cases, nurses and doctors need to provide service together. This assumption might lead to a different model.

33 The Combination of Patient-Call Treatments and Nurse-Initiated Treatments

In reality, patients that do not "call" a nurse need to be checked and treated every fixed period of time. This means that if the patient has not called the nurse for T units of time, he will join the queue anyway, since the nurse will initiate the treatment. The consequence is that the time between inter-arrivals is not exponential but truncated-exponential, and the nurse station is of type G/M/s. There are other modeling possibilities, such as using a multi-class chain in which one can divide the flow out of the nurse node into two classes, each with different exponential dormant parameters, as was done in the case of heterogeneous patients above (see Section 31).

34 Two Service Stations

Suppose one has two stations of the M/M/S type with s_1 and s_2 servers (denoted nodes 1 and 2 respectively), one entrance node of the type M/M/1, and all other nodes are M/M/∞ nodes. We believe that the appropriate QED conditions will be:

$$\lim_{\lambda\to\infty} \frac{n - s_1 - s_2 - A}{\sqrt{A}} = \eta, \qquad -\infty < \eta < \infty; \tag{i}$$

$$\lim_{\lambda\to\infty} \sqrt{s_1}\left(1 - \frac{\rho_1}{s_1}\right) = \beta_1, \qquad -\infty < \beta_1 < \infty; \tag{ii}$$

$$\lim_{\lambda\to\infty} \sqrt{s_2}\left(1 - \frac{\rho_2}{s_2}\right) = \beta_2, \qquad -\infty < \beta_2 < \infty. \tag{iii}$$

But there are also other possibilities, such as working with one node in the QED regime, and the other node in the ED regime (see Section 32).

Appendices

A Appendices of Part I

A.1 Steady State Measures: Proof

Proof of Theorem 1 in Section 4.

Proof. Proof of the probability of waiting (α) is based on the Arrival Theorem for open Jackson networks [17]. Thus, based on the steady-state distribution, using striate forward calculation we obtain:

$$\alpha = P\left(Q_1(\infty) \geq s\right) = \sum_{j=0}^{\infty}\sum_{i=s}^{\infty} P_{ij} = \sum_{j=0}^{\infty}\sum_{i=s}^{\infty} \frac{(R_1)^i}{\nu(i)}\pi_{01}\frac{(R_2)^j}{j!}\pi_{02} = \left[\frac{(R_1)^s}{s!(1-R_1/s)}\right]\pi_{01}.$$

Defining the marginal distribution P_i as the probability to have i customers in node 1, it follows that:

$$P_i = \sum_j P_{ij} = \frac{(R_1)^i}{\nu(i)}\pi_{01}.$$

If L_q is defined as the number of customers in queue for service, and $\rho = \frac{R_1}{s}$, then:

$$P\left(L_q = i\right) = P_{i+s} = \frac{(R_1)^{i+s}}{s!s^i}\pi_{01} = \alpha\left(1-\frac{R_1}{s}\right)\left(\frac{R_1}{s}\right)^i = \alpha\left(1-\rho\right)\left(\rho\right)^i.$$

If a customer becomes Needy when there are already i other Needy customers in the system, he will need to wait an in-queue random waiting time that follows an Erlang distribution with $(i-s+1)^+$ stages, each with rate $s\mu$. The probability that this Erlang-distributed random variable is greater than t is $\int_t^\infty \frac{s\mu(s\mu x)^{i-s}}{(i-s)!}$. Clearly, the patient only waits if $i > s$. Defining E_k as a random variable with the following Erlang distribution: $E_k \sim Erlang(k, s\mu)$, it follows that $P\left(W > t\right)$ can be found in the following way:

$$\begin{aligned}
&P\left(W > t\right) = \sum_{j=0}^{\infty}\sum_{i=s}^{\infty} P_{ij}P(E_{i-s+1} > t) = \sum_{i=s}^{\infty} P_i \int_t^\infty \frac{\mu s(\mu s x)^{i-s}}{(i-s)!}e^{-\mu s x}dx \\
&= \sum_{i=0}^{\infty} P_{i+s}\int_t^\infty \frac{\mu s(\mu s x)^{i}}{i!}e^{-\mu s x}dx = \sum_{i=0}^{\infty}\alpha\left(1-\rho\right)\rho^i\int_t^\infty \frac{\mu s(\mu s x)^{i}}{i!}e^{-\mu s x}dx \\
&= \alpha\mu s\left(1-\rho\right)\int_t^\infty e^{-\mu s x}\sum_{i=0}^{\infty}\frac{(\mu R_1 x)^i}{i!}dx = \alpha\mu s\left(1-\rho\right)\int_t^\infty e^{-\mu s x}e^{\mu R_1 x}dx \\
&= \alpha\mu s\left(1-\rho\right)\int_t^\infty e^{-\mu s(1-\rho)x}dx = \alpha e^{-s\mu(1-\rho)t}.
\end{aligned} \tag{A.1}$$

Using $P\left(W>t\right)$ we can define the expected waiting of customer $E[W]$ as

$$E[W]=\int_0^\infty P\left(W>t\right)dt=\int_0^\infty \alpha e^{-s\mu(1-\rho)t}dt=\frac{\alpha}{\mu s(1-\rho)}.$$

□

A.2 The Offered Load Measure: Proofs

Proof of Theorem 2 in Section 5.1.

Proof. If S_i are exponentially distributed, the Erlang-R model is actually a 2-node state-dependent stochastic network, denoted as $(M_t/M/S_t^k)^K$ where $S_t^k\in\{1,2,...,\infty\}$, $k\in 1,2$ and $K=2$. We examine a corresponding queueing network $(M_t/M/\infty)^2$ with the same structure, service rates, and arrival rate. The only difference is that the number of servers is infinite in every node.

Let $Q^\infty=\{Q^\infty(t),t\geq 0\}$ be a 2-dimensional stochastic queueing process, where $Q^\infty(t)=(Q_1^\infty(t),Q_2^\infty(t))$: $Q_1^\infty(t)$ representing the number of *Needy* patients in the system, and $Q_2^\infty(t)$ the number of *Content* patients in the system, at time t.

The process $Q^\infty(t)$ satisfies the following equations:

$$\begin{aligned}
Q_1^\infty(t)&=Q_1^\infty(0)+A_1^a\left(\int_0^t\lambda_u du\right)-A_2^d\left(\int_0^t p\mu Q_1^\infty(u)du\right)-A_{12}\left(\int_0^t(1-p)\mu Q_1^\infty(u)du\right)\\
&\quad+A_{21}\left(\int_0^t\delta Q_2(u)du\right)\\
Q_2^\infty(t)&=Q_2^\infty(0)+A_{12}\left(\int_0^t p\mu Q_1^\infty(u)du\right)-A_{21}\left(\int_0^t\delta Q_2^\infty(u)du\right),
\end{aligned}$$

where A_1^a,A_2^d,A_{12} and A_{21} are four mutually independent, standard (mean rate 1), Poisson processes. We now introduce a family of scaled queues, indexed by $\eta>0$, so that both the arrival rate and the number of nurses grow together to infinity, i.e. scaled up by η, but leave the Needy and Content rates unscaled:

$$\begin{aligned}
Q_1^{\eta,\infty}(t)&=Q_1^{\eta,\infty}(0)+A_1^a\left(\int_0^t\eta\lambda_u du\right)-A_2^d\left(\int_0^t p\mu Q_1^{\eta,\infty}(u)du\right)\\
&\quad-A_{12}\left(\int_0^t(1-p)\mu Q_1^{\eta,\infty}(u)du\right)+A_{21}\left(\int_0^t\delta Q_2^{\eta,\infty}(u)du\right)\\
&=Q_1^{\eta,\infty}(0)+A_1^a\left(\int_0^t\eta\lambda_u du\right)-A_2^d\left(\int_0^t\eta p\mu\frac{1}{\eta}Q_1^{\eta,\infty}(u)du\right)\\
&\quad-A_{12}\left(\int_0^t\eta(1-p)\mu\frac{1}{\eta}Q_1^{\eta,\infty}(u)du\right)+A_{21}\left(\int_0^t\eta\delta\left(\frac{1}{\eta}Q_2^{\eta,\infty}(u)\right)du\right),\\
Q_2^{\eta,\infty}(t)&=Q_2^{\eta,\infty}(0)+A_{12}\left(\int_0^t p\mu Q_1^{\eta,\infty}(u)du\right)-A_{21}\left(\int_0^t\delta Q_2^{\eta,\infty}(u)du\right)\\
&=Q_2^{\eta,\infty}(0)+A_{12}\left(\int_0^t\eta p\mu\frac{1}{\eta}Q_1^{\eta,\infty}(u)du\right)-A_{21}\left(\int_0^t\eta\delta\left(\frac{1}{\eta}Q_2^{\eta,\infty}(u)\right)du\right).
\end{aligned}$$

By Theorem 2.2 (FSLLN) in [53],

$$\lim_{\eta\to\infty}\frac{Q^{\eta,\infty}(t)}{\eta}=Q^{(0)}(t)\quad a.s.,$$

where $Q^{(0)}(t)$ is called the *fluid approximation* and is the solution of the following ODE:

$$Q_1^{(0),\infty}(t)=Q_1^{(0),\infty}(0)+\int_0^t\left(\lambda_u-\mu Q_1^{(0),\infty}(u)+\delta Q_2^{(0),\infty}(u)\right)du$$
$$Q_2^{(0),\infty}(t)=Q_2^{(0),\infty}(0)+\int_0^t\left(p\mu Q_1^{(0),\infty}(u)-\delta Q_2^{(0),\infty}(u)\right)du.$$

Note that $R(t)=Q^{(0),\infty}(t)$ by definition. Therefore, (5.2) is actually the solution for the following ODE:

$$\frac{d}{dt}R_1(t)=\lambda_t+\delta R_2(t)-\mu R_1(t),$$
$$\frac{d}{dt}R_2(t)=p\mu R_1(t)-\delta R_2(t).$$

□

Proof of Proposition 1 in Section 5.2.

Proof. The proof of Proposition 1 can be derived from straightforward calculations:

$$\begin{aligned}
R_1(t)&=\sum_{j=0}^{\infty}p^jE[S_1]E\left[\lambda(t-S_{1,e}-S_1^{*j}-S_2^{*j})\right]\\
&=\sum_{j=0}^{\infty}p^jE[S_1]E\left[a+b(t-S_{1,e}-S_1^{*j}-S_2^{*j})\right]\\
&=\frac{a+bt}{1-p}E[S_1]-\sum_{j=0}^{\infty}p^jE[S_1]E\left[b(S_{1,e}+S_1^{*j}+S_2^{*j})\right]\\
&=\frac{a+bt}{1-p}E[S_1]-E[S_1]b\sum_{j=0}^{\infty}p^jE\left[S_{1,e}\right]-E[S_1]b\sum_{j=0}^{\infty}p^jE\left[S_1^{*j}\right]-E[S_1]b\sum_{j=0}^{\infty}p^jE\left[S_2^{*j}\right]\\
&=\frac{a+bt}{1-p}E[S_1]-E[S_1]bE\left[S_{1,e}\right]\sum_{j=0}^{\infty}p^j-E[S_1]bE\left[S_1\right]\sum_{j=0}^{\infty}jp^j-E[S_1]bE\left[S_2\right]\sum_{j=0}^{\infty}jp^j\\
&=\frac{a+bt}{1-p}E[S_1]-E[S_1]bE\left[S_{1,e}\right]\frac{1}{1-p}-E[S_1]bE\left[S_1\right]\frac{p}{(1-p)^2}-E[S_1]bE\left[S_2\right]\frac{p}{(1-p)^2}\\
&=\lambda(t-E\left[S_{1,e}\right])\frac{E[S_1]}{1-p}-b\frac{p}{(1-p)^2}E[S_1]\left(E\left[S_1\right]+E\left[S_2\right]\right).
\end{aligned}$$

□

Proof of Theorem 3 in Section 5.2.

Proof. The proof of Theorem 3 can be derived from straightforward calculations:

$$
\begin{aligned}
R_1(t) &= \sum_{j=0}^{\infty} p^j E[S_1] E\left[\lambda(t - S_{1,e} - S_1^{*j} - S_2^{*j})\right] \\
&= \sum_{j=0}^{\infty} p^j E[S_1] E\left[\lambda(t) - \lambda^{(1)}(t)\left(S_{1,e} + S_1^{*j} + S_2^{*j}\right) + \frac{1}{2}\lambda^{(2)}(t)\left(S_{1,e} + S_1^{*j} + S_2^{*j}\right)^2\right] \\
&= \sum_{j=0}^{\infty} p^j E[S_1] E\left[\lambda(t) - \lambda^{(1)}(t)\left(S_{1,e} + S_1^{*j} + S_2^{*j}\right)\right] \\
&+ \frac{1}{2}\lambda^{(2)}(t) E[S_1] \sum_{j=0}^{\infty} p^j E\left[\left(S_{1,e} + S_1^{*j} + S_2^{*j}\right)^2\right] \\
&= \frac{E[S_1]}{1-p}\left[\lambda(t) - \lambda^{(1)}(t)\left(E\left[S_{1,e}\right] + \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)\right] \\
&+ \frac{1}{2}\lambda^{(2)}(t) E[S_1] \sum_{j=0}^{\infty} p^j E\left[S_{1,e}^2 + (S_1^{*j})^2 + (S_2^{*j})^2 + 2S_{1,e}S_2^{*j} + 2S_1^{*j}S_2^{*j} + 2S_{1,e}S_1^{*j}\right] \\
&= \frac{E[S_1]}{1-p}\left[\lambda(t) - \lambda^{(1)}(t)\left(E\left[S_{1,e}\right] + \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)\right] + \frac{1}{2}\lambda^{(2)}(t) E[S_1] \sum_{j=0}^{\infty} p^j \left(E[S_{1,e}^2]\right. \\
&\left. + jVAR[S_1] + j^2 E[S_1]^2 + jVAR[S_2] + j^2 E[S_2]^2 + 2jE[S_{1,e}]E[S_2] + 2j^2 E[S_1]E[S_2] + 2jE[S_{1,e}]E[S_1]\right) \\
&= \frac{E[S_1]}{1-p}\left[\lambda(t) - \lambda^{(1)}(t)\left(E\left[S_{1,e}\right] + \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)\right] + \frac{1}{2}\lambda^{(2)}(t) E[S_1]\left(\frac{1}{1-p}VAR[S_{1,e}]\right. \\
&+ \frac{1}{1-p}E[S_{1,e}]^2 + \frac{p}{(1-p)^2}\left(VAR[S_1] + VAR[S_2] + 2E[S_{1,e}]E[S_2] + 2E[S_{1,e}]E[S_1]\right) \\
&\left. + \frac{p(1+p)}{(1-p)^3}\left(E[S_1]^2 + E[S_2]^2 + 2E[S_1]E[S_2]\right)\right) \\
&= \frac{E[S_1]}{1-p}\left[\lambda(t) - \lambda^{(1)}(t)\left(E\left[S_{1,e}\right] + \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)\right. \\
&+ \frac{1}{2}\lambda^{(2)}(t)\left(E[S_{1,e}]^2 + 2\frac{p}{1-p}E[S_{1,e}]\left(E[S_1] + E[S_2]\right) + \frac{p^2}{(1-p)^2}\left(E[S_1] + E[S_2]\right)^2\right) \\
&\left. + \frac{1}{2}\lambda^{(2)}(t)\left(VAR[S_{1,e}] + \frac{p}{1-p}(VAR[S_1] + VAR[S_2]) + \frac{p}{(1-p)^2}\left(E[S_1] + E[S_2]\right)^2\right)\right] \\
&= \frac{E[S_1]}{1-p}\left[\lambda(t) - \lambda^{(1)}(t)\left(E\left[S_{1,e}\right] + \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)\right. \\
&+ \frac{1}{2}\lambda^{(2)}(t)\left(E\left[S_{1,e}\right] + \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)^2 \\
&\left. + \frac{1}{2}\lambda^{(2)}(t)\left(VAR[S_{1,e}] + \frac{p}{1-p}(VAR[S_1] + VAR[S_2]) + \frac{p}{(1-p)^2}\left(E[S_1] + E[S_2]\right)^2\right)\right] \\
&= \frac{E[S_1]}{1-p}\left[\lambda\left(t - E\left[S_{1,e}\right] - \frac{p}{1-p}\left(E\left[S_1\right] + E\left[S_2\right]\right)\right)\right. \\
&\left. + \frac{1}{2}\lambda^{(2)}(t)\left(VAR[S_{1,e}] + \frac{p}{1-p}(VAR[S_1] + VAR[S_2]) + \frac{p}{(1-p)^2}\left(E[S_1] + E[S_2]\right)^2\right)\right].
\end{aligned}
$$

□

A.3 The Offered-Load for Sinusoidal Arrival Rate: Proofs

Proof of Theorem 4 in Section 5.3.1.

Proof. Assuming S_i is exponentially distributed, $S_{i,e} = S_i$, knowing that a sum of j identical exponential variables, with rate μ, has an Erlang distribution with parameters μ and j, we can find explicit expressions for $E[\cos(\omega(S_1^{*j_1} + S_2^{*j_2}))]$, and $E[\sin(\omega(S_1^{*j_1} + S_2^{*j_2}))]$. Defining $x \equiv S_1^{*j_1} \sim Erlang(\mu, j_1)$, and $y \equiv S_2^{*j_2} \sim Erlang(\delta, j_2)$:

$$
\begin{aligned}
&E[\cos(\omega(S_1^{*j_1} + S_2^{*j_2}))] = E[\cos(\omega(x+y))] \\
&= \int_0^\infty \int_0^\infty \frac{1}{2}\left(e^{i\omega(x+y)} + e^{-i\omega(x+y)}\right) \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} \frac{\delta^{j_2} y^{j_2-1} e^{-\delta y}}{(j_2-1)!} dxdy = \\
&= \frac{1}{2}\int_0^\infty \int_0^\infty e^{i\omega x} e^{i\omega y} \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} \frac{\delta^{j_2} y^{j_2-1} e^{-\delta y}}{(j_2-1)!} dxdy \\
&+ \frac{1}{2}\int_0^\infty \int_0^\infty e^{-i\omega x} e^{-i\omega y} \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} \frac{\delta^{j_2} y^{j_2-1} e^{-\delta y}}{(j_2-1)!} dxdy \\
&= \frac{1}{2}\int_0^\infty e^{i\omega x} \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} dx \int_0^\infty e^{i\omega y} \frac{\delta^{j_2} y^{j_2-1} e^{-\delta y}}{(j_2-1)!} dy \\
&+ \frac{1}{2}\int_0^\infty e^{-i\omega x} \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} dx \int_0^\infty e^{-i\omega y} \frac{\delta^{j_2} y^{j_2-1} e^{-\delta y}}{(j_2-1)!} dy \\
&= \frac{1}{2}\int_0^\infty \frac{\mu^{j_1} x^{j_1-1} e^{-(\mu-i\omega)x}}{(j_1-1)!} dx \int_0^\infty \frac{\delta^{j_2} y^{j_2-1} e^{-(\delta-i\omega)y}}{(j_2-1)!} dy \\
&+ \frac{1}{2}\int_0^\infty \frac{\mu^{j_1} x^{j_1-1} e^{-(\mu+i\omega)x}}{(j_1-1)!} dx \int_0^\infty \frac{\delta^{j_2} y^{j_2-1} e^{-(\delta+i\omega)y}}{(j_2-1)!} dy \\
&= \frac{1}{2}\frac{\mu^{j_1}}{(\mu-i\omega)^{j_1}} \int_0^\infty \frac{(\mu-i\omega)^{j_1} x^{j_1-1} e^{-(\mu-i\omega)x}}{(j_1-1)!} dx \frac{\delta^{j_2}}{(\delta-i\omega)^{j_2}} \int_0^\infty \frac{(\delta-i\omega)^{j_2} y^{j_2-1} e^{-(\delta-i\omega)y}}{(j_2-1)!} dy \\
&+ \frac{1}{2}\frac{\mu^{j_1}}{(\mu+i\omega)^{j_1}} \int_0^\infty \frac{(\mu+i\omega)^{j_1} x^{j_1-1} e^{-(\mu+i\omega)x}}{(j_1-1)!} dx \frac{\delta^{j_2}}{(\delta+i\omega)^{j_2}} \int_0^\infty \frac{(\delta+i\omega)^{j_2} y^{j_2-1} e^{-(\delta+i\omega)y}}{(j_2-1)!} dy \\
&= \frac{1}{2}\frac{\mu^{j_1}}{(\mu-i\omega)^{j_1}} \frac{\delta^{j_2}}{(\delta-i\omega)^{j_2}} + \frac{1}{2}\frac{\mu^{j_1}}{(\mu+i\omega)^{j_1}} \frac{\delta^{j_2}}{(\delta+i\omega)^{j_2}} \\
&= \frac{1}{2}\mu^{j_1}\delta^{j_2}\left(\frac{1}{(\mu-i\omega)^{j_1}(\delta-i\omega)^{j_2}} + \frac{1}{(\mu+i\omega)^{j_1}(\delta+i\omega)^{j_2}}\right)
\end{aligned}
\tag{A.2}
$$

and

$$
\begin{aligned}
&E[\sin(\omega(S_1^{*j_1} + S_2^{*j_2}))] = E[\sin(\omega(x+y))] \\
&= \int_0^\infty \int_0^\infty \frac{1}{2i}\left(e^{i\omega(x+y)} - e^{-i\omega(x+y)}\right) \frac{\mu^{j_1} x^{j_1-1} e^{-\mu x}}{(j_1-1)!} \frac{\delta^{j_2} y^{j_2-1} e^{-\delta y}}{(j_2-1)!} dxdy
\end{aligned}
\tag{A.3}
$$

$$
\begin{aligned}
&= \frac{1}{2i}\int_0^\infty\int_0^\infty e^{i\omega x}e^{i\omega y}\frac{\mu^{j_1}x^{j_1-1}e^{-\mu x}}{(j_1-1)!}\frac{\delta^{j_2}y^{j_2-1}e^{-\delta y}}{(j_2-1)!}dxdy\\
&- \frac{1}{2i}\int_0^\infty\int_0^\infty e^{-i\omega x}e^{-i\omega y}\frac{\mu^{j_1}x^{j_1-1}e^{-\mu x}}{(j_1-1)!}\frac{\delta^{j_2}y^{j_2-1}e^{-\delta y}}{(j_2-1)!}dxdy\\
&= \frac{1}{2i}\int_0^\infty e^{i\omega x}\frac{\mu^{j_1}x^{j_1-1}e^{-\mu x}}{(j_1-1)!}dx\int_0^\infty e^{i\omega y}\frac{\delta^{j_2}y^{j_2-1}e^{-\delta y}}{(j_2-1)!}dy\\
&- \frac{1}{2i}\int_0^\infty e^{-i\omega x}\frac{\mu^{j_1}x^{j_1-1}e^{-\mu x}}{(j_1-1)!}dx\int_0^\infty e^{-i\omega y}\frac{\delta^{j_2}y^{j_2-1}e^{-\delta y}}{(j_2-1)!}dy\\
&= \frac{1}{2i}\int_0^\infty \frac{\mu^{j_1}x^{j_1-1}e^{-(\mu-i\omega)x}}{(j_1-1)!}dx\int_0^\infty \frac{\delta^{j_2}y^{j_2-1}e^{-(\delta-i\omega)y}}{(j_2-1)!}dy\\
&- \frac{1}{2i}\int_0^\infty \frac{\mu^{j_1}x^{j_1-1}e^{-(\mu+i\omega)x}}{(j_1-1)!}dx\int_0^\infty \frac{\delta^{j_2}y^{j_2-1}e^{-(\delta+i\omega)y}}{(j_2-1)!}dy\\
&= \frac{1}{2i}\frac{\mu^{j_1}}{(\mu-i\omega)^{j_1}}\int_0^\infty \frac{(\mu-i\omega)^{j_1}x^{j_1-1}e^{-(\mu-i\omega)x}}{(j_1-1)!}dx\frac{\delta^{j_2}}{(\delta-i\omega)^{j_2}}\int_0^\infty \frac{(\delta-i\omega)^{j_2}y^{j_2-1}e^{-(\delta-i\omega)y}}{(j_2-1)!}dy\\
&- \frac{1}{2i}\frac{\mu^{j_1}}{(\mu+i\omega)^{j_1}}\int_0^\infty \frac{(\mu+i\omega)^{j_1}x^{j_1-1}e^{-(\mu+i\omega)x}}{(j_1-1)!}dx\frac{\delta^{j_2}}{(\delta+i\omega)^{j_2}}\int_0^\infty \frac{(\delta+i\omega)^{j_2}y^{j_2-1}e^{-(\delta+i\omega)y}}{(j_2-1)!}dy\\
&= \frac{1}{2i}\frac{\mu^{j_1}}{(\mu-i\omega)^{j_1}}\frac{\delta^{j_2}}{(\delta-i\omega)^{j_2}} - \frac{1}{2i}\frac{\mu^{j_1}}{(\mu+i\omega)^{j_1}}\frac{\delta^{j_2}}{(\delta+i\omega)^{j_2}}\\
&= \frac{1}{2i}\mu^{j_1}\delta^{j_2}\left(\frac{1}{(\mu-i\omega)^{j_1}(\delta-i\omega)^{j_2}} - \frac{1}{(\mu+i\omega)^{j_1}(\delta+i\omega)^{j_2}}\right).
\end{aligned}
$$

Incorporating (A.2) and (A.3) into (5.10) yields:

$$
\begin{aligned}
&R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + E[S_1]\bar{\lambda}\kappa\sum_{j=0}^\infty p^j E\left[\sin(\omega t)\cos(\omega(S_1^{*j+1}+S_2^{*j})) - \sin(\omega(S_1^{*j+1}+S_2^{*j}))\cos(\omega t)\right] =\\
&= \frac{E[S_1]\bar{\lambda}}{1-p} + E[S_1]\bar{\lambda}\kappa\sum_{j=0}^\infty p^j\left[\sin(\omega t)\frac{1}{2}\mu^{(j+1)}\delta^j\left(\frac{1}{(\mu-i\omega)^{(j+1)}(\delta-i\omega)^j} + \frac{1}{(\mu+i\omega)^{(j+1)}(\delta+i\omega)^j}\right)\right.\\
&\left. - \cos(\omega t)\frac{1}{2i}\mu^{(j+1)}\delta^j\left(\frac{1}{(\mu-i\omega)^{(j+1)}(\delta-i\omega)^j} - \frac{1}{(\mu+i\omega)^{(j+1)}(\delta+i\omega)^j}\right)\right]\\
&= \frac{E[S_1]\bar{\lambda}}{1-p} + E[S_1]\bar{\lambda}\kappa\frac{1}{2}\sum_{j=0}^\infty(p\mu\delta)^j\mu\left[\sin(\omega t)\left(\frac{1}{(\mu-i\omega)^{j+1}(\delta-i\omega)^j} + \frac{1}{(\mu+i\omega)^{j+1}(\delta+i\omega)^j}\right)\right.\\
&\left. - \cos(\omega t)\frac{1}{i}\left(\frac{1}{(\mu-i\omega)^{j+1}(\delta-i\omega)^j} - \frac{1}{(\mu+i\omega)^{j+1}(\delta+i\omega)^j}\right)\right]\\
&= \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\sum_{j=0}^\infty(p\mu\delta)^j\frac{1}{(\mu-i\omega)^{j+1}(\delta-i\omega)^j} + \frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\sum_{j=0}^\infty(p\mu\delta)^j\frac{1}{(\mu+i\omega)^{j+1}(\delta+i\omega)^j}\\
&- \frac{1}{2}\bar{\lambda}\kappa\cos(\omega t)\frac{1}{i}\sum_{j=0}^\infty(p\mu\delta)^j\frac{1}{(\mu-i\omega)^{j+1}(\delta-i\omega)^j} + \frac{1}{2}\bar{\lambda}\kappa\cos(\omega t)\frac{1}{i}\sum_{j=0}^\infty(p\mu\delta)^j\frac{1}{(\mu+i\omega)^{j+1}(\delta+i\omega)^j}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa \sin(\omega t)\frac{1}{(\mu - i\omega)} \sum_{j=0}^{\infty} \left(\frac{p\mu\delta}{(\mu - i\omega)(\delta - i\omega)}\right)^j \\
&+ \frac{1}{2}\bar{\lambda}\kappa \sin(\omega t)\frac{1}{(\mu + i\omega)} \sum_{j=0}^{\infty} \left(\frac{p\mu\delta}{(\mu + i\omega)(\delta + i\omega)}\right)^j - \frac{1}{2}\bar{\lambda}\kappa \cos(\omega t)\frac{1}{i(\mu - i\omega)} \sum_{j=0}^{\infty} \left(\frac{p\mu\delta}{(\mu - i\omega)(\delta - i\omega)}\right)^j \\
&+ \frac{1}{2}\bar{\lambda}\kappa \cos(\omega t)\frac{1}{i(\mu + i\omega)} \sum_{j=0}^{\infty} \left(\frac{p\mu\delta}{(\mu + i\omega)(\delta + i\omega)}\right)^j \\
&= \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa \sin(\omega t)\frac{1}{(\mu - i\omega)} \frac{(\mu - i\omega)(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} + \frac{1}{2}\bar{\lambda}\kappa \sin(\omega t)\frac{1}{(\mu + i\omega)} \frac{(\mu + i\omega)(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta} \\
&- \frac{1}{2}\bar{\lambda}\kappa \cos(\omega t)\frac{1}{i(\mu - i\omega)} \frac{(\mu - i\omega)(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} + \frac{1}{2}\bar{\lambda}\kappa \cos(\omega t)\frac{1}{i(\mu + i\omega)} \frac{(\mu + i\omega)(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta} \\
&= \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa \sin(\omega t)\left[\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} + \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}\right] \\
&- \frac{1}{2i}\bar{\lambda}\kappa \cos(\omega t)\left[\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} - \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}\right] \\
&= \frac{E[S_1]\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sqrt{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}} \cos(\omega t + \pi + tan^{-1}(\theta)),
\end{aligned}
$$

where

$$
\theta = i \cdot \frac{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} + \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}{\frac{(\delta - i\omega)}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} - \frac{(\delta + i\omega)}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}} = \frac{-\mu(-\delta^2 + p\delta^2 - \omega^2)}{\omega(\delta^2 + \omega^2 + p\mu\delta)}.
$$

□

Proof of Theorem 5 in Section 5.3.1.

Proof. Incorporating (A.2) and (A.3) into (5.11) yields:

$$
\begin{aligned}
\lambda_1^+(t) &= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j E\left[\sin(\omega t)\cos(\omega(S_1^{*j} + S_2^{*j})) - \sin(\omega(S_1^{*j} + S_2^{*j}))\cos(\omega t)\right] = \\
&= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j \left[\sin(\omega t)\frac{1}{2}\mu^j\delta^j\left(\frac{1}{(\mu - i\omega)^j(\delta - i\omega)^j} + \frac{1}{(\mu + i\omega)^j(\delta + i\omega)^j}\right)\right. \\
&\left. - \cos(\omega t)\frac{1}{2i}\mu^j\delta^j\left(\frac{1}{(\mu - i\omega)^j(\delta - i\omega)^j} - \frac{1}{(\mu + i\omega)^j(\delta + i\omega)^j}\right)\right] \\
&= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\frac{1}{2} \sum_{j=0}^{\infty} (p\mu\delta)^j \left[\sin(\omega t)\left(\frac{1}{(\mu - i\omega)^j(\delta - i\omega)^j} + \frac{1}{(\mu + i\omega)^j(\delta + i\omega)^j}\right)\right. \\
&\left. - \cos(\omega t)\frac{1}{i}\left(\frac{1}{(\mu - i\omega)^j(\delta - i\omega)^j} - \frac{1}{(\mu + i\omega)^j(\delta + i\omega)^j}\right)\right] \\
&= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\frac{1}{2}\sin(\omega t)\left[\sum_{j=0}^{\infty}\left(\frac{p\mu\delta}{(\mu - i\omega)(\delta - i\omega)}\right)^j + \sum_{j=0}^{\infty}\left(\frac{p\mu\delta}{(\mu + i\omega)(\delta + i\omega)}\right)^j\right] \\
&- \bar{\lambda}\kappa\frac{1}{2i}\cos(\omega t)\left[\sum_{j=0}^{\infty}\left(\frac{p\mu\delta}{(\mu - i\omega)(\delta - i\omega)}\right)^j - \sum_{j=0}^{\infty}\left(\frac{p\mu\delta}{(\mu + i\omega)(\delta + i\omega)}\right)^j\right]
\end{aligned}
$$

$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\frac{1}{2}\sin(\omega t)\left[\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} + \frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$- \bar{\lambda}\kappa\frac{1}{2i}\cos(\omega t)\left[\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} - \frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\sqrt{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}\cdot\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}\cos(\omega t+\pi+tan^{-1}(\theta))$$

where

$$\theta = i\cdot\frac{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} + \frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}}{\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} - \frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}} = \frac{\omega^2\delta^2+\omega^4+\omega^2p\mu\delta+\mu^2\delta^2-\mu^2p\delta^2+\mu^2\omega^2}{\mu\omega p\delta(\mu+\delta)}$$

□

Proof of Proposition 2 in Section 5.3.1.

Proof. The limits are obtained from Proposition 4 by straightforward calculations. Starting with limits as a function of ω:

$$\lim_{\omega\to 0} R_1(t) = \frac{E[S_1]\bar{\lambda}}{1-p} + \lim_{\omega\to 0}\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} + \frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$- \frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} - \frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$= \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa\lim_{\omega\to 0}\sin(\omega t)\left[\frac{\delta}{\mu\delta-p\mu\delta} + \frac{\delta}{\mu\delta-p\mu\delta}\right]$$
$$- \frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{\delta}{\mu\delta-p\mu\delta} - \frac{\delta}{\mu\delta-p\mu\delta}\right]$$
$$= \frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa\lim_{\omega\to 0}\sin(\omega t)\left[\frac{2\delta}{\mu\delta(1-p)}\right] - 0$$
$$= \lim_{\omega\to 0}\frac{E[S_1]\bar{\lambda}}{1-p} + E[S_1]\frac{\bar{\lambda}}{\mu(1-p)}\kappa\sin(\omega t)$$

$$\lim_{\omega\to\infty} R_1(t) = \lim_{\omega\to\infty}\frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} + \frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$- \frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} - \frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$= \lim_{\omega\to\infty}\frac{E[S_1]\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)[0+0] - \frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)[0-0] = \frac{E[S_1]\bar{\lambda}}{1-p}$$

and

$$\lim_{\omega\to 0}\lambda_1^+(t) = \lim_{\omega\to 0}\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\frac{1}{2}\sin(\omega t)\left[\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} + \frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$- \bar{\lambda}\kappa\frac{1}{2i}\cos(\omega t)\left[\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta} - \frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]$$
$$= \lim_{\omega\to 0}\frac{\bar{\lambda}}{1-p} + \frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{2\delta\mu}{\delta\mu(1-p)}\right] - 0$$
$$= \lim_{\omega\to 0}\frac{\bar{\lambda}}{1-p} + \frac{\bar{\lambda}}{1-p}\kappa\sin(\omega t)$$

$$\begin{aligned}
\lim_{\omega\to\infty}\lambda_1^+(t) &= \lim_{\omega\to\infty}\frac{\bar{\lambda}}{1-p}+\bar{\lambda}\kappa\frac{1}{2}\sin(\omega t)\left[\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}+\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&-\bar{\lambda}\kappa\frac{1}{2i}\cos(\omega t)\left[\frac{(\mu-i\omega)(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}-\frac{(\mu+i\omega)(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&=\lim_{\omega\to\infty}\frac{\bar{\lambda}}{1-p}+\bar{\lambda}\kappa\frac{1}{2}\sin(\omega t)[1+1]-\bar{\lambda}\kappa\frac{1}{2i}\cos(\omega t)[1-1]\\
&=\lim_{\omega\to\infty}\frac{\bar{\lambda}}{1-p}+\bar{\lambda}\kappa\sin(\omega t)
\end{aligned}$$

The limits as a function of δ are proven as follows:

$$\begin{aligned}
\lim_{\delta\to 0}R_1(t) &= \frac{E[S_1]\bar{\lambda}}{1-p}+\lim_{\delta\to 0}\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}+\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}-\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{-i\omega}{(\mu-i\omega)(-i\omega)}+\frac{i\omega}{(\mu+i\omega)(i\omega)}\right]\\
&-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{-i\omega}{(\mu-i\omega)(-i\omega)}-\frac{i\omega}{(\mu+i\omega)(i\omega)}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{1}{\mu-i\omega}+\frac{1}{\mu+i\omega}\right]-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{1}{\mu-i\omega}-\frac{1}{\mu+i\omega}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{2\mu}{(\mu-i\omega)(\mu+i\omega)}\right]-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{2i\omega}{(\mu-i\omega)(\mu+i\omega)}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{\bar{\lambda}\kappa}{\mu^2+\omega^2}\left(\mu\sin(\omega t)-\omega\cos(\omega t)\right)
\end{aligned}$$

The extreme values of $R_1(t)$ in this case are $\max_t(R_1(t))=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{\bar{\lambda}\kappa}{\sqrt{\mu^2+\omega^2}}$. Thus, the relative amplitude is $\frac{1}{\sqrt{\mu^2+\omega^2}}$. As $\delta\to\infty$ we obtain:

$$\begin{aligned}
\lim_{\delta\to\infty}R_1(t) &= \frac{E[S_1]\bar{\lambda}}{1-p}+\lim_{\delta\to\infty}\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}+\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{(\delta-i\omega)}{(\mu-i\omega)(\delta-i\omega)-p\mu\delta}-\frac{(\delta+i\omega)}{(\mu+i\omega)(\delta+i\omega)-p\mu\delta}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{1}{(1-p)\mu-i\omega}+\frac{1}{(1-p)\mu+i\omega}\right]\\
&-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{1}{(1-p)\mu-i\omega}-\frac{1}{(1-p)\mu+i\omega}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{1}{2}\bar{\lambda}\kappa\sin(\omega t)\left[\frac{2(1-p)\mu}{((1-p)\mu-i\omega)((1-p)\mu+i\omega)}\right]\\
&-\frac{1}{2i}\bar{\lambda}\kappa\cos(\omega t)\left[\frac{2i\omega}{((1-p)\mu-i\omega)((1-p)\mu+i\omega)}\right]\\
&=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{\bar{\lambda}\kappa}{(1-p)^2\mu^2+\omega^2}\left((1-p)\mu\sin(\omega t)-\omega\cos(\omega t)\right)
\end{aligned}$$

The extreme values of $R_1(t)$ in this case are $\max_t(R_1(t))=\frac{E[S_1]\bar{\lambda}}{1-p}+\frac{\bar{\lambda}\kappa}{\sqrt{(1-p)^2\mu^2+\omega^2}}$. Thus, the relative amplitude is $\frac{1}{\sqrt{(1-p)^2\mu^2+\omega^2}}$. □

A.4 Comparison to Erlang-C: Proof

Proof of Theorem 6 in Section 5.3.2.

Proof. We need to prove that $AmpRatio < 1$. $AmpRatio$ is given by:

$$AmpRatio = \sqrt{\frac{\delta^2+\omega^2}{((\mu-i\omega)(\delta-i\omega)-p\mu\delta)((\mu+i\omega)(\delta+i\omega)-p\mu\delta)}} \Big/ \frac{1}{\sqrt{((1-p)\mu)^2+\omega^2}}$$

Thus we need to prove that:

$$\frac{(\delta^2+\omega^2)((1-p)^2\mu^2+\omega^2)}{[(\mu-i\omega)(\delta-i\omega)-p\mu\delta][(\mu+i\omega)(\delta+i\omega)-p\mu\delta]} \overset{?}{<} 1$$

$$\frac{\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2+\delta^2\omega^2+\omega^4}{(\mu-i\omega)(\delta-i\omega)(\mu+i\omega)(\delta+i\omega)-p\mu\delta[(\mu+i\omega)(\delta+i\omega)+(\mu-i\omega)(\delta-i\omega)]+p^2\mu^2\delta^2} \overset{?}{<} 1$$

$$\frac{\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2+\delta^2\omega^2+\omega^4}{(\mu^2+\omega^2)(\delta^2+\omega^2)-p\mu\delta(2\mu\delta-2\omega^2)+p^2\mu^2\delta^2} \overset{?}{<} 1$$

$$\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2+\delta^2\omega^2+\omega^4 \overset{?}{<} \mu^2\delta^2+\omega^2\delta^2+\mu^2\omega^2+\omega^4+2p\mu\delta(\omega^2-\mu\delta)+p^2\mu^2\delta^2$$

$$\delta^2(1-p)^2\mu^2+\omega^2(1-p)^2\mu^2 \overset{?}{<} \mu^2\omega^2+\mu^2\delta^2(1-p)^2+2p\mu\delta\omega^2$$

$$\omega^2(1-p)^2\mu^2 \overset{?}{<} \mu^2\omega^2+2p\mu\delta\omega^2$$

Which is true for every μ, δ, p, and ω, since $0 \le p \le 1$.

In the second part of the theorem, we need to prove that $AmpRatio$ reaches its minimum when $\omega = \sqrt{\delta\mu(1-p)}$. The derivative of $AmpRatio$ by ω is:

$$\frac{dAmpRatio}{d\omega} = \frac{2p\omega\mu(2\delta+(2-p)\mu)(\omega^2+(1-p)\mu\delta)(\omega^2-(1-p)\mu\delta)}{(\omega^4+(p-1)^2\delta^2\mu^2+\omega^2(\delta^2+2p\delta\mu+\mu^2))^2}$$

This derivative equals zero when $\omega = 0$ or $\omega = \sqrt{\delta\mu(1-p)}$. When $\omega = 0$ the $AmpRatio$ reaches its maximum which is 1, and when $\omega = \sqrt{\delta\mu(1-p)}$ it reaches its minimum value. □

A.5 Analysis of the Cases: Sinusoidal Arrival Rates and Deterministic Service Times

We will now analyze $R_1(t)$ (5.10), and $\lambda_1^+(t)$ (5.11) for the case of deterministic service times. In this case it is simpler to develop an expression for $\lambda_1^+(t)$, and on its basis the expression of $R_1(t)$.

Theorem 22. *If S_i are deterministic then,*

$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{\frac{e^{i(\omega t-\frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right\}$$

and

$$R_1(t) = S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\left[Re\left\{\frac{\frac{1}{-i\omega}(e^{i(\omega(t-S_1)-\frac{\pi}{2})}-e^{i(\omega t-\frac{\pi}{2})})}{1-pe^{-i\omega(S_1+S_2)}}\right\}\right].$$

Proof. We will begin with $\lambda_1^+(t)$. In the deterministic case the following holds: $E[S_i^{*j}] = jS_i$. Consequently,

$$\lambda_1^+(t) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j E[\sin(\omega(t - S_1^{*j} + S_2^{*j}))] = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j \sin(\omega(t - jS_1 + jS_2))$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j \cos\left(\omega(t - j(S_1+S_2)) - \frac{\pi}{2}\right) = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \sum_{j=0}^{\infty} p^j Re\left\{e^{i(\omega(t-j(S_1+S_2))-\frac{\pi}{2})}\right\}$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{\sum_{j=0}^{\infty} p^j e^{i(\omega t - \frac{\pi}{2})} e^{-ij\omega(S_1+S_2)}\right\} = \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{e^{i(\omega t-\frac{\pi}{2})} \sum_{j=0}^{\infty} p^j e^{-ij\omega(S_1+S_2)}\right\}$$
$$= \frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa Re\left\{\frac{e^{i(\omega t-\frac{\pi}{2})}}{1 - pe^{-i\omega(S_1+S_2)}}\right\}$$

In order to find an expression for $R_1(t)$, we use the fact that when $S_i \sim D$, $S_{i,e}$ is Uniformly distributed $[0, D]$. Therefore:

$$R_1(t) = E[S_1]E[\lambda^+(t - S_{1,e})] = S_1\frac{\bar{\lambda}}{1-p} + S_1\bar{\lambda}\kappa E\left[Re\left\{\frac{e^{i(\omega(t-S_{1,e})-\frac{\pi}{2})}}{1 - pe^{-i\omega(S_1+S_2)}}\right\}\right]$$
$$= S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa \int_0^{S_1}\left[Re\left\{\frac{e^{i(\omega(t-x)-\frac{\pi}{2})}}{1 - pe^{-i\omega(S_1+S_2)}}\right\}\right] dx = S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\left[Re\left\{\frac{\int_0^{S_1} e^{i(\omega(t-x)-\frac{\pi}{2})}dx}{1 - pe^{-i\omega(S_1+S_2)}}\right\}\right]$$
$$= S_1\frac{\bar{\lambda}}{1-p} + \bar{\lambda}\kappa\left[Re\left\{\frac{\frac{1}{-i\omega}(e^{i(\omega(t-S_1)-\frac{\pi}{2})} - e^{i(\omega t-\frac{\pi}{2})})}{1 - pe^{-i\omega(S_1+S_2)}}\right\}\right]$$

□

The amplitude of $\lambda_1^+(t)$ is given by: $\left|\bar{\lambda}\kappa\frac{e^{i(\omega t-\frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right|$. The shape of this function with respect to $\omega(S_1+S_2)$ is shown in Figure 91. Since the sine function is bounded between $[-1, 1]$, the maximum amplitude of $\lambda_1^+(t)$ will be achieved when $\omega j(S_1+S_2) = 2\pi$ for all j, since j is an integer. This means that $\omega(S_1+S_2) = 2\pi$. In this case the returning streams from node 2 are fully synchronized with the external input stream ($\lambda(t)$), and the infinite sum will converge to $\frac{1}{1-p}\sin\omega t$. Therefore, in this case the relative amplitude (i.e. the relation between the amplitude of $\lambda_1^+(t)$ to $\lambda(t)$) is $\frac{1}{1-p}$, as seen in Figure 91. On the other end, the minimal amplitude of $\lambda_1^+(t)$ will be achieved when $\omega(S_1+S_2) = \pi$. In this case the returning streams from node 2 will be balancing the external input stream ($\lambda(t)$), and the relative amplitude will be $\frac{1}{1+p}$. Other interesting points are when $\omega(S_1+S_2) = \pi/2$ and $\omega(S_1+S_2) = 3/2\pi$. In these cases, the amplitude is $\frac{1}{1+p^2}\sin\omega t$, i.e. the relative amplitude is $\frac{1}{1+p^2}$. In general when $0 < \omega(S_1+S_2) < \pi$ the amplitude is decreasing with respect to the sum S_1+S_2, and when $\pi < \omega(S_1+S_2) < 2\pi$ the amplitude is increasing.

Note that due to the shape of the amplitude function, special care is required when optimizing the system. If one seeks to shorten the length of stay of customers in the system, he can usually

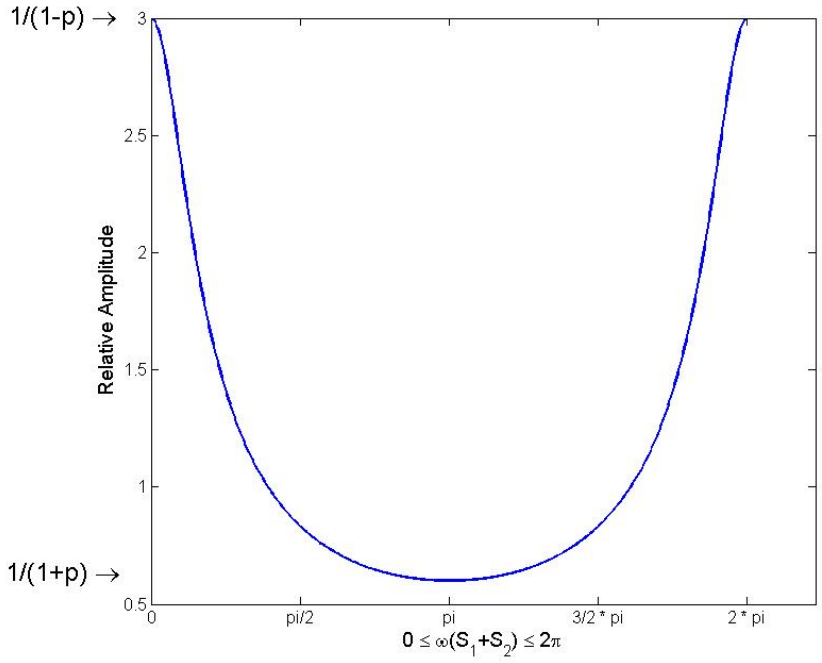


Figure 91: Plot of relative amplitude of $\lambda_1^+(t)$ with respect to ω

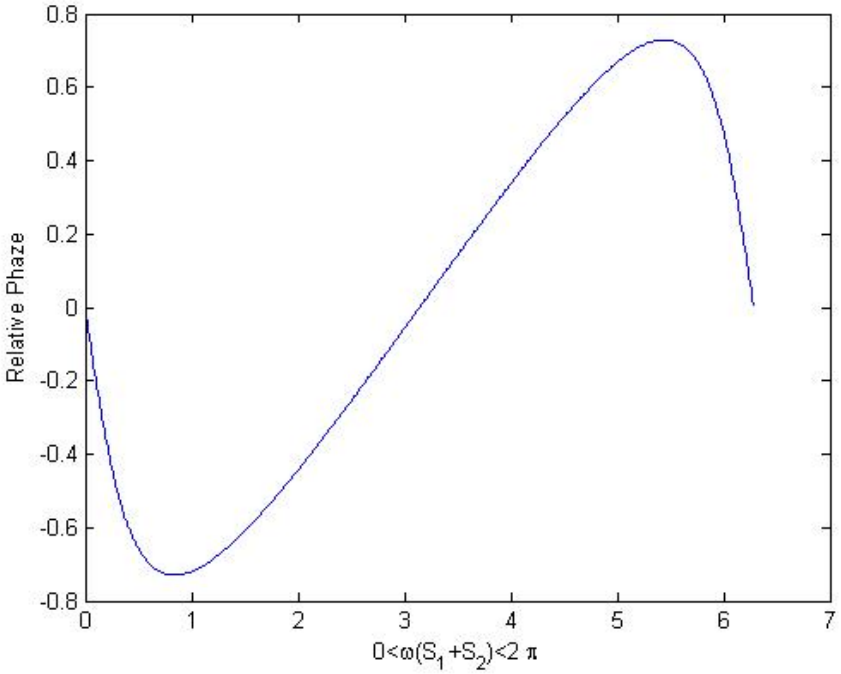


Figure 92: Plot of relative phase of $\lambda_1^+(t)$ with respect to ω

influence the Needy and Content times (S_1 and S_2 respectively). However, if the system operates in the decreasing region, shortening S_1 or S_2 will increase the amplitude of $\lambda_1^+(t)$, and therefore the amplitude of $R_1(t)$ will also increase. Staffing in a system in which the amplitude is large and staffing demands are changing rapidly over time, is much more difficult to operate than a system with a small amplitude in which the staffing is more stable over time. This is especially true in small systems.

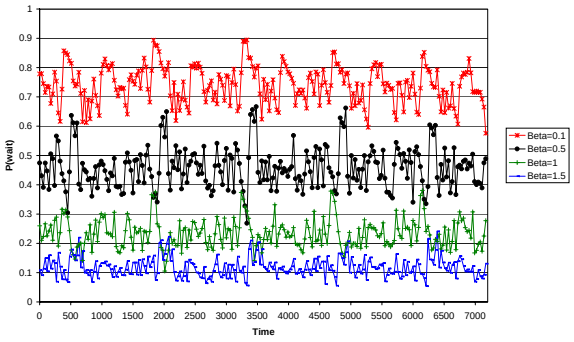
The phase shift of $\lambda_1^+(t)$ is given by: $Angle\left(\frac{e^{i(\omega t-\frac{\pi}{2})}}{1-pe^{-i\omega(S_1+S_2)}}\right)$. The shape of this function with respect to ω is shown in Figure 92. The phase of the aggregated arrivals is determined by a combination of arrival rates of returning customers with different phases. We will analyze it (denote as t_s), with respect to $\omega(S_1+S_2)$. If the arrival streams are fully synchronized, i.e. $\omega(S_1+S_2)=2\pi$ or $\omega(S_1+S_2)=\pi$, then $t_s=0$. The maximum phase shift depends on p, and its value is 0.7297.

A.6 Validation of MOL Staffing: More Examples in Case Study 2

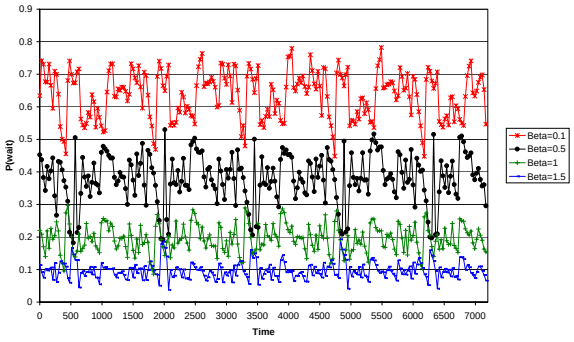
Since the rounding has a very large impact in small systems, we also tried to round the staffing levels in a different way: If $s(t)$ is less than 1, round up to 1; if $s(t)$ is greater than 1, round down values that are less than +0.2, otherwise round up. In this rounding procedure the actual β was closer to the predefined one (when compared to the round up procedure), though the main difference was in small values of β. Figure 93 shows the performance measure $P(W>0)$ over time for various values of β. We see that the performance measure is relatively stable, and that the four scenarios are separable. We can compare the two using the MSE (Mean Square Error) measure. Table 6 shows the MSE rates for each β, for the measures $P(W>0)$ and $E[W]$. While for the $P(W>0)$ measure no significant difference is observed between the two, for $E[W]$ simple roundup works much better.

Measure	Beta	Round Special	RoundUP
$P(W>0)$	0.1	0.98	1.27
	0.5	1.22	1.39
	1	0.59	0.41
	1.5	0.28	0.16
E[W]	0.1	13105.24	2899.99
	0.5	754.05	100.70
	1	45.69	13.63
	1.5	9.52	4.29

Table 6: MSE measure for two rounding procedures



(special rounding)



(roundup)

Figure 93: Case study 2: Simulation results of $P(W > 0)$ for various β values in small systems

B Approximating the Number of Needy Customers and Waiting Times Using Fluid and Diffusion Limits

In this section, we develop Fluid and Diffusion limits for Erlang-R. We then use these approximations to analyze mass-casualty events in which the arrival rate changes rapidly during a short time. While it is clear that fluid approximations are very useful in analyzing time-varying systems, these approximations are also useful in understanding the transient behavior of a system in steady-state. For example, we might need to evaluate the probability that the number of customers (patients) in the system will exceed a certain threshold during a certain time horizon. This is useful when setting control rules for the EW, for example starting special procedures such as ambulance diversion and calling for additional staffing. The answer to such questions requires diffusion approximations, such as the ones we develop here.

The Erlang-R model is a state-dependent open queueing network. We follow the mathematical framework of Mandelbaum et al. [53] on time-varying queues. This framework give us a general solution, suitable for time-varying arrivals, and time-varying staffing policies. Note that with general time-varying arrival rates ($\lambda(t),\ t \geq 0$), the ODE system we develop here is unlikely to be tractable analytically. Nevertheless, we can solve it numerically.

The Erlang-R model is, in fact, a 2-node state-dependent stochastic network denoted as $(M_t/M/S_t^k)^K$ where $S_t^k \in \{1, 2, ..., \infty\},\ k \in 1, 2$ and $K = 2$. Let $Q = \{Q(t), t \geq 0\}$ be a 2-dimensional stochastic queueing process, where $Q(t) = (Q_1(t), Q_2(t))$: $Q_1(t)$ representing the number of *Needy* patients in the system (i.e., those either waiting for service or being served), and $Q_2(t)$ the number of *Content* patients in the system, at time t. The process $Q(t)$ satisfies the following equations:

$$
\begin{aligned}
Q_1(t) &= Q_1(0) + A_1^a\left(\int_0^t \lambda_u du\right) - A_2^d\left(\int_0^t p\mu\left(Q_1(u) \wedge s_u\right) du\right) - A_{12}\left(\int_0^t (1-p)\mu\left(Q_1(u) \wedge s_u\right) du\right) \\
&\quad + A_{21}\left(\int_0^t \delta Q_2(u) du\right) \\
Q_2(t) &= Q_2(0) + A_{12}\left(\int_0^t p\mu\left(Q_1(u) \wedge s_u\right) du\right) - A_{21}\left(\int_0^t \delta Q_2(u) du\right),
\end{aligned}
$$

where A_1^a, A_2^d, A_{12} and A_{21} are four mutually independent, standard (mean rate 1), Poisson processes. We now introduce a family of scaled queues, indexed by $\eta > 0$, so that both the arrival rate and the number of nurses grow together to infinity, i.e. scaled up by η, but leave the Needy and Content

rates unscaled:

$$\begin{aligned}
Q_1^\eta(t) &= Q_1^\eta(0) + A_1^a\left(\int_0^t \eta\lambda_u du\right) - A_2^d\left(\int_0^t p\mu\left(Q_1^\eta(u)\wedge \eta s_u\right)du\right) \\
&\quad - A_{12}\left(\int_0^t (1-p)\mu\left(Q_1^\eta(u)\wedge \eta s_u\right)du\right) + A_{21}\left(\int_0^t \delta Q_2^\eta(u)du\right) \\
&= Q_1^\eta(0) + A_1^a\left(\int_0^t \eta\lambda_u du\right) - A_2^d\left(\int_0^t \eta p\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right) \\
&\quad - A_{12}\left(\int_0^t \eta(1-p)\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right) + A_{21}\left(\int_0^t \eta\delta\left(\frac{1}{\eta}Q_2^\eta(u)\right)du\right), \\
Q_2^\eta(t) &= Q_2^\eta(0) + A_{12}\left(\int_0^t p\mu\left(Q_1^\eta(u)\wedge \eta s_u\right)du\right) - A_{21}\left(\int_0^t \delta Q_2^\eta(u)du\right) \\
&= Q_2^\eta(0) + A_{12}\left(\int_0^t \eta p\mu\left(\frac{1}{\eta}Q_1^\eta(u)\wedge s_u\right)du\right) - A_{21}\left(\int_0^t \eta\delta\left(\frac{1}{\eta}Q_2^\eta(u)\right)du\right).
\end{aligned} \tag{B.1}$$

Theorem 23. *(FSLLN) Using the scaling of* (B.1)*, we have*

$$\lim_{\eta\to\infty}\frac{Q^\eta(t)}{\eta} = Q^{(0)}(t) \quad a.s.,$$

where $Q^{(0)}(t)$ *is called the* fluid approximation *and is the solution of the following ODE:*

$$\begin{aligned}
Q_1^{(0)}(t) &= Q_1^{(0)}(0) + \int_0^t \left(\lambda_u - \mu\left(Q_1^{(0)}(u)\wedge s_u\right) + \delta Q_2^{(0)}(u)\right)du \\
Q_2^{(0)}(t) &= Q_2^{(0)}(0) + \int_0^t \left(p\mu\left(Q_1^{(0)}(u)\wedge s_u\right) - \delta Q_2^{(0)}(u)\right)du.
\end{aligned} \tag{B.2}$$

This is based on Theorem 2.2 in [53]. Equation (B.2) is equivalent to the following representation

$$\begin{cases}
\frac{dQ_1^{(0)}}{dt}(t) &= \lambda_t + \delta Q_2^{(0)}(t) - \mu(s_t \wedge Q_1^{(0)}(t)) \\
\frac{dQ_2^{(0)}}{dt}(t) &= -\delta Q_2^{(0)}(t) + p\mu(s_t \wedge Q_1^{(0)}(t))
\end{cases} \tag{B.3}$$

We continue by developing the diffusion limits of the Erlang-R model. These diffusion limits will be used to develop variance and covariance phrases that enable us to develop statistical boundaries for the number of patients in the system. The fluid and diffusion processes can be used in order to analyze mass-casualty events as well as other time-varying scenarios, as demonstrated in Section 8.

Theorem 24. *(FCLT) Using the scaling of* (B.1)*, and the fluid limits* (B.3) *we have*

$$\lim_{\eta\to\infty}\sqrt{\eta}\left[\frac{Q^\eta(t)}{\eta} - Q^{(0)}(t)\right] \stackrel{d}{=} Q^{(1)}(t), \tag{B.4}$$

where $Q^{(1)}(t)$ *is called the* diffusion approximation *and is the solution of the following SDE (Stochas-*

tic Differential Equation):

$$
\begin{aligned}
Q_1^{(1)}(t) &= Q_1^{(1)}(0) + \int_0^t \left(\mu 1_{\left\{Q_1^{(0)}(u) \le s_u\right\}} Q_1^{(1)}(u)^- - \mu 1_{\left\{Q_1^{(0)}(u) < s_u\right\}} Q_1^{(1)}(u)^+ + \delta Q_2^{(1)}(u) \right) du \\
&\quad + B_1^a \left(\int_0^t \lambda_u du \right) - B_2^d \left(\int_0^t p\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) - B_{12} \left(\int_0^t (1-p)\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) \\
&\quad + B_{21} \left(\int_0^t \delta Q_2^{(0)}(u) du \right), \\
Q_2^{(1)}(t) &= Q_2^{(1)}(0) + \int_0^t \left(p\mu 1_{\left\{Q_1^{(0)}(u) < s_u\right\}} Q_1^{(1)}(u)^+ - p\mu 1_{\left\{Q_1^{(0)}(u) \le s_u\right\}} Q_1^{(1)}(u)^- - \delta Q_2^{(1)}(u) \right) du \\
&\quad + B_{12} \left(\int_0^t p\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) - B_{21} \left(\int_0^t \delta Q_2^{(0)}(u) du \right),
\end{aligned}
\tag{B.5}
$$

where B_1^a, B_2^d, B_{12} and B_{21} are four mutually independent, standard (mean is 0 and the variance at time t is t) Brownian motions; $x^+ \equiv \max(x,0)$, and $x^- \equiv \max(-x,0) = -\min(x,0)$.

This is based on Theorem 2.3 in [53].

The following theorem presents the mean vector and the covariance matrix for the diffusion limit.

Theorem 25. *Using the scaling of (B.1), the mean vector for the diffusion limit (B.5) solves the following DE:*

$$
\begin{aligned}
\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)\right] &= \mu 1_{\left\{Q_1^{(0)}(t) \le s_t\right\}} \mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \mu 1_{\left\{Q_1^{(0)}(t) < s_t\right\}} \mathrm{E}\left[Q_1^{(1)}(t)^+\right] + \delta \mathrm{E}\left[Q_2^{(1)}(t)\right], \\
\frac{d}{dt}\mathrm{E}\left[Q_2^{(1)}(t)\right] &= p\mu 1_{\left\{Q_1^{(0)}(t) < s_t\right\}} \mathrm{E}\left[Q_1^{(1)}(t)^+\right] - p\mu 1_{\left\{Q_1^{(0)}(t) \le s_t\right\}} \mathrm{E}\left[Q_1^{(1)}(t)^-\right] \\
&\quad - \delta \mathrm{E}\left[Q_2^{(1)}(t)\right].
\end{aligned}
\tag{B.6}
$$

and the covariance matrix for the diffusion limit solves the following DE:

$$
\begin{aligned}
\frac{d}{dt}\mathrm{Var}\left[Q_1^{(1)}(t)\right] &= 2\mu 1_{\left\{Q_1^{(0)}(t) \le s_t\right\}} \mathrm{Cov}\left[Q_1^{(1)}(t), Q_1^{(1)}(t)^-\right] - 2\mu 1_{\left\{Q_1^{(0)}(t) < s_t\right\}} \mathrm{Cov}\left[Q_1^{(1)}(t), Q_1^{(1)}(t)^+\right] \\
&\quad + 2\delta \mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] + \lambda_t + \mu\left(Q_1^{(0)}(t) \wedge s_t\right) + \delta Q_2^{(0)}(t), \\
\frac{d}{dt}\mathrm{Var}\left[Q_2^{(1)}(t)\right] &= 2p\mu 1_{\left\{Q_1^{(0)}(t) < s_t\right\}} \mathrm{Cov}\left[Q_2^{(1)}(t), Q_1^{(1)}(t)^+\right] \\
&\quad - 2p\mu 1_{\left\{Q_1^{(0)}(t) \le s_t\right\}} \mathrm{Cov}\left[Q_2^{(1)}(t), Q_1^{(1)}(t)^-\right] - 2\delta \mathrm{Var}\left[Q_2^{(1)}(t)\right] \\
&\quad + p\mu\left(Q_1^{(0)}(t) \wedge s_t\right) + \delta Q_2^{(0)}(t), \\
\frac{d}{dt}\mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] &= \mu 1_{\left\{Q_1^{(0)}(t) \le s_t\right\}} \mathrm{Cov}\left[Q_2^{(1)}(t), Q_1^{(1)}(t)^-\right] - \mu 1_{\left\{Q_1^{(0)}(t) < s_t\right\}} \mathrm{Cov}\left[Q_2^{(1)}(t), Q_1^{(1)}(t)^+\right] \\
&\quad + \delta\left(\mathrm{Var}\left[Q_2^{(1)}(t)\right] - \mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right]\right) + p\mu 1_{\left\{Q_1^{(0)}(t) < s_t\right\}} \mathrm{Cov}\left[Q_1^{(1)}(t), Q_1^{(1)}(t)^+\right] \\
&\quad - p\mu 1_{\left\{Q_1^{(0)}(t) \le s_t\right\}} \mathrm{Cov}\left[Q_1^{(1)}(t), Q_1^{(1)}(t)^-\right] - \delta Q_2^{(0)}(t) - p\mu\left(Q_1^{(0)}(t) \wedge s_t\right).
\end{aligned}
$$

(B.7)

B.1 Essentially Negligible Critical Regime and Applications to the Analysis of Mass-Casualty Events

The differential equations (B.6) - (B.7) are not easy to solve. Therefore, we will also assume that the time the system spends in critically-loaded values is negligible. Formally, define S as the set of times when the system is critically loaded, i.e., the times when the number of nurses equals the number of customers in the Needy state:

$$S = \{t > 0 | Q_1^{(0)}(t) = s_t\}.$$

We assume that the set S has measure zero, which practically means that the process $Q_1^{(1)}(t)$ passes through the state with equal numbers of nurses and patients in Needy state very quickly. This assumption is reasonable, as for a normally distributed process the measure of a single point is zero. In this case:

$$\mu 1_{\left\{Q_1^{(0)}(u) < s_u\right\}} Q_1^{(1)}(u)^+ - \mu 1_{\left\{Q_1^{(0)}(u) \leq s_u\right\}} Q_1^{(1)}(u)^- = \mu 1_{\left\{Q_1^{(0)}(u) \leq s_u\right\}} Q_1^{(1)}(u)$$

Proposition 4. *Assume that the set of time points S has measure zero. Then (B.5) becomes:*

$$\begin{aligned}
Q_1^{(1)}(t) &= Q_1^{(1)}(0) + \int_0^t \left(-\mu 1_{\left\{Q_1^{(0)}(u) \leq s_u\right\}} Q_1^{(1)}(u) + \delta Q_2^{(1)}(u) \right) du + B_1^a \left(\int_0^t \lambda_u du \right) \\
&\quad - B_2^d \left(\int_0^t p\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) - B_{12} \left(\int_0^t (1-p)\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) \\
&\quad + B_{21} \left(\int_0^t \delta Q_2^{(0)}(u) du \right), \qquad \text{(B.8)} \\
Q_2^{(1)}(t) &= Q_2^{(1)}(0) + \int_0^t \left(p\mu 1_{\left\{Q_1^{(0)}(u) \leq s_u\right\}} Q_1^{(1)}(u) - \delta Q_2^{(1)}(u) \right) du \\
&\quad + B_{12} \left(\int_0^t p\mu \left(Q_1^{(0)}(u) \wedge s_u \right) du \right) - B_{21} \left(\int_0^t \delta Q_2^{(0)}(u) du \right),
\end{aligned}$$

The mean vector for the diffusion approximation (B.6) is then:

$$\begin{aligned}
\frac{d}{dt} \mathrm{E}\left[Q_1^{(1)}(t) \right] &= -\mu 1_{\left\{Q_1^{(0)}(t) \leq s_t\right\}} \mathrm{E}\left[Q_1^{(1)}(t) \right] + \delta \mathrm{E}\left[Q_2^{(1)}(t) \right], \\
\frac{d}{dt} \mathrm{E}\left[Q_2^{(1)}(t) \right] &= p\mu 1_{\left\{Q_1^{(0)}(t) \leq s_t\right\}} \mathrm{E}\left[Q_1^{(1)}(t) \right] - \delta \mathrm{E}\left[Q_2^{(1)}(t) \right],
\end{aligned}$$

and the variance matrix (B.7) *is:*

$$\begin{aligned}
\frac{d}{dt}\mathrm{Var}\left[Q_1^{(1)}(t)\right] &= -2\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Var}\left[Q_1^{(1)}(t)\right] + 2\delta\mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] \\
&\quad + \lambda_t + \mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t), \\
\frac{d}{dt}\mathrm{Var}\left[Q_2^{(1)}(t)\right] &= -2\delta\mathrm{Var}\left[Q_2^{(1)}(t)\right] + 2p\mu\mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] \\
&\quad + p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t), \\
\frac{d}{dt}\mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] &= -\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}} + \delta\right)\mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] + \delta\mathrm{Var}\left[Q_2^{(1)}(t)\right] \\
&\quad + p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{Var}\left[Q_1^{(1)}(t)\right] - p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) - \delta Q_2^{(0)}(t).
\end{aligned} \tag{B.9}$$

B.1.1 A Numerical Example

We preformed a simulation of mass-casualty events. When such an event is in progress, the EW must, over a short time period, take care of the regular patients, release the ones who can be released, and give emergency care to the new patients. We can examine the effect of such an event on the EW, and the time it takes to overcome such an emergency situation, using the models developed in the previous section.

We demonstrate a simulation vs. fluid and diffusion approximation. We calculate upper and lower envelopes using the following expression: $Q_i^{(0)}(t) \pm \sqrt{\mathrm{Var}\left[Q_i^{(1)}(t)\right]}$, for $i = 1, 2$, where $Q_i^{(0)}(t)$ is the numerical solution of the ODE (B.2), and $\mathrm{Var}\left[Q_i^{(1)}(t)\right]$ is the numerical solution of the ODE (B.9).

In this example, we set the parameters as follows:

$$\lambda_t = \begin{cases} 50 & \text{if } 9 \le t \le 11, \\ 10 & \text{otherwise,} \end{cases}$$

$\delta = 0.2,\ \mu = 1,\ p = 0.25$, and $s_t = 50$.

The left diagram of Figure 94 shows the number of customers in the system over time. It shows the fluid solution and the simulation results. $Q1$ is the number of Needy customers, and $Q2$ is the number of content customers. The right diagram of Figure 94 shows the changes in the average number of Needy customers ($Q_1^{(0)}$) and upper and lower envelops of this process. We see that both fluid and diffusion approximations are remarkably accurate in this case. This is a case of essentially negligible critical regime, in which the strong approximations work well. This event illustrates a situation in which, during a very short period of two hours (9-11), the arrival rate is multiplied fivefold. As a consequence, the number of patients in the medical unit is also multiplied by five,

although the peak of the load is reached only towards the end of the peak period at 11. From that point, the number of patients gradually decreases to normal levels. It takes several more hours for the system to stabilize again.

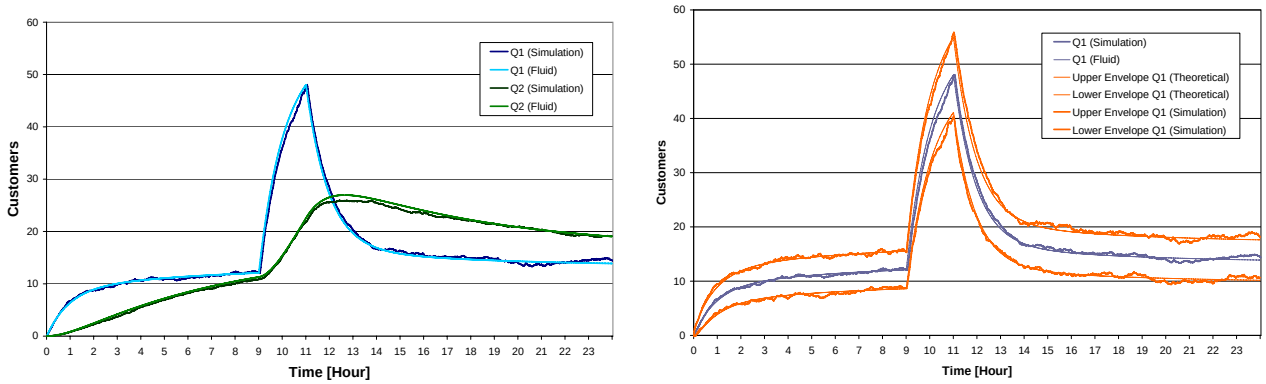


Figure 94: Numerical example 1: Mass arrival at interval (9,11)

B.2 The Virtual Waiting Time Process

In this section we develop an approximation to the process of virtual waiting time. The virtual waiting time is the in-queue waiting time for a hypothetical patient, who just became Needy. We rely on 24, 23 and the corollary of Puhalskii [65] to prove the following:

Theorem 26. *Using the scaling of* (B.1)*, the virtual waiting time process is given by:*

$$\sqrt{\eta}(W^{\eta}(t)) \xrightarrow{d} \left[\frac{Q_1^{(1)}(0) + V_1(t) - U_1(Z(t))}{\mu\left(Q_1^{(0)}(Z(t)) \wedge s_{Z(t)}\right)} \right]^{+}.$$

B.2.1 Numerical Examples

In order to examine our approximation for the process $W(t)$, the virtual waiting time, we used the following examples, taken from Mandelbaum et al. [55]. These use a periodic arrival rate with two values: low and hight. The first example data is:

$$\lambda_t = \begin{cases} 20 & \text{if } t \in \{[0,4),[6,9),[11,14),[16,24]\}, \\ 100 & \text{otherwise,} \end{cases}$$

$\delta = 0.2,\ \mu = 1,\ p = 0.25$, and $s_t = 70$.

Figure 95 shows changes in the number of customers in each node. It compares the simulation results compared to the fluid and diffusion approximation. As before, we see remarkable matching

between the two. Figure 96 shows the changes in $E[W]$ over time. Again, we compare fluid approximations to the simulated approximations. We observe that there is a good match between the two. Differences are observed in underloaded times, when the fluid approximates no waiting while simulation observe positive waiting.

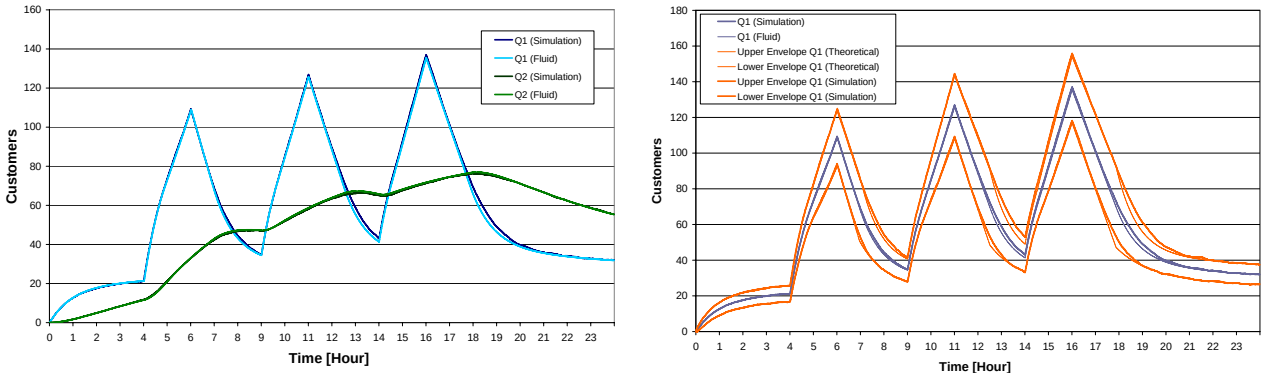


Figure 95: Numerical example 3: Fluid approximation vs. simulation results of Q(t)

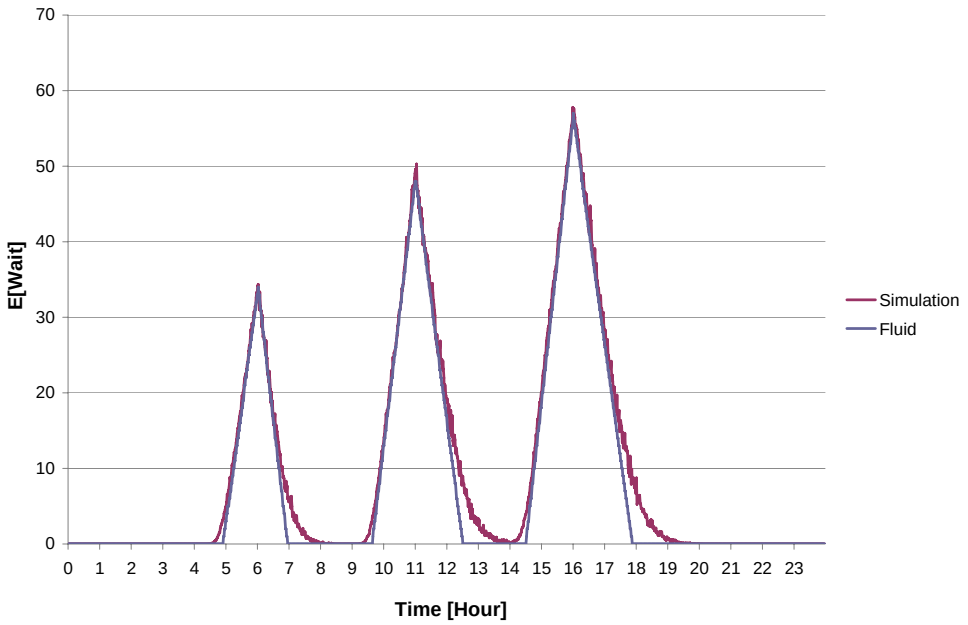


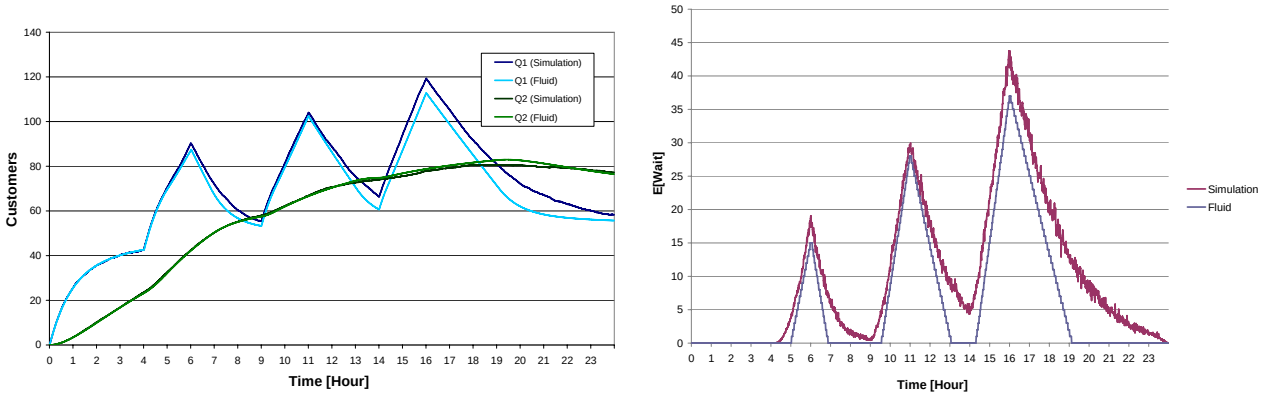
Figure 96: Numerical example 3: Fluid approximation vs. simulation results of E[W(t)]

The second example data is:

$$\lambda_t = \begin{cases} 40 & \text{if } t \in \{[0,4),[6,9),[11,14),[16,24]\}, \\ 80 & \text{otherwise,} \end{cases}$$

$\delta = 0.2$, $\mu = 1$, $p = 0.25$, and $s_t = 70$. The left diagram of Figure 97 shows changes in the number of customers in each node. It compares the simulation results to the fluid and diffusion

approximation. We see that in this example $Q_2(t)$ simulation matches the fluid, but $Q_1(t)$ has lower match. The right diagram of Figure 97 shows the changes in $E[W]$ over time. Again, we compare fluid approximations to the simulated ones. We observe that there is a difference between the two, although the shape of the two is similar.



$Q(t)$ $E[W(t)]$

Figure 97: Numerical example 4: Fluid approximation vs. simulation results

B.3 Fluid and Diffusion Limits for the Number of Needy Customers: Proof

Proof of Theorem 25 in Section B.

Proof. The mean vector is based on Theorem 2.4 in [53]. We now prove (B.7). The variance and covariance processes are given by:

$$\begin{aligned}
&\mathrm{Var}\left[Q_1^{(1)}(t)\right] = \mathrm{E}\left[(Q_1^{(1)}(t))^2\right] - \left(\mathrm{E}\left[Q_1^{(1)}(t)\right]\right)^2, \\
&\mathrm{Var}\left[Q_2^{(1)}(t)\right] = \mathrm{E}\left[(Q_2^{(1)}(t))^2\right] - \left(\mathrm{E}\left[Q_2^{(1)}(t)\right]\right)^2, \\
&\mathrm{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] = \mathrm{E}\left[Q_1^{(1)}(t)Q_2^{(1)}(t)\right] - \mathrm{E}\left[Q_1^{(1)}(t)\right]\mathrm{E}\left[Q_2^{(1)}(t)\right].
\end{aligned}$$

We must determine the derivative of the expressions $\frac{d}{dt}\mathrm{E}[X_tY_t]$ and $\frac{d}{dt}\mathrm{E}[X_t]\,\mathrm{E}[Y_t]$, where $(X_t)_{t\in\mathbb{R}_0^+}$ and $(Y_t)_{t\in\mathbb{R}_0^+}$ are time-dependent stochastic processes. By the chain rule:

$$\frac{d}{dt}\left(\mathrm{E}[X_t]\,\mathrm{E}[Y_t]\right) = \frac{d}{dt}\mathrm{E}[X_t]\cdot\mathrm{E}[Y_t] + \mathrm{E}[X_t]\cdot\frac{d}{dt}\mathrm{E}[Y_t]$$

and by the chain rule of stochastic calculus (Ito formula):

$$d\left(\mathrm{E}[X_tY_t]\right) = \mathrm{E}[dX_t\cdot Y_t] + \mathrm{E}[X_t\cdot dY_t] + \mathrm{E}[dX_t\cdot dY_t].$$

Note that by the assumption of Brownian Motion (BM): $dB_i(t)dB_j(t) = \delta_{ij}(t)dt$, $dt \cdot dt = dt \cdot dB(t) = dB(t) \cdot dt = 0$, where $B_i(t)$ and $B_j(t)$ are mutually independent standard BMs and δ_{ij} denote the Kronecker-Delta, i.e.

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

We now show the values of $\frac{d}{dt}\mathrm{E}\left[(Q_1^{(1)}(t))^2\right]$, $\frac{d}{dt}\mathrm{E}\left[(Q_2^{(1)}(t))^2\right]$,and $\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)Q_2^{(1)}(t)\right]$. Using (B.5) we obtain:

$$\begin{aligned}
dQ_1^{(1)}(t) &= \left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ + \delta Q_2^{(1)}(t)\right)dt \\
&\quad + \sqrt{\lambda_t}dB_1^a(t) - \sqrt{p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_2^d(t) - \sqrt{(1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_{12}(t) \\
&\quad + \sqrt{\delta Q_2^{(0)}(t)}dB_{21}(t), \\
dQ_2^{(1)}(t) &= \left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \delta Q_2^{(1)}(t)\right)dt \\
&\quad + \sqrt{(1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_{12}(t) - \sqrt{\delta Q_2^{(0)}(t)}dB_{21}(t).
\end{aligned}$$

$$\begin{aligned}
d\left(Q_1^{(1)}(t)\right)^2 &= 2\cdot dQ_1^{(1)}(t)\cdot Q_1^{(1)}(t) + dQ_1^{(1)}(t)\cdot dQ_1^{(1)}(t) \\
&= 2Q_1^{(1)}(t)\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ + \delta Q_2^{(1)}(t)\right)dt \\
&\quad + 2Q_1^{(1)}(t)\sqrt{\lambda_t}dB_1^a(t) - 2Q_1^{(1)}(t)\sqrt{p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_2^d(t) \\
&\quad - 2Q_1^{(1)}(t)\sqrt{(1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_{12}(t) + 2Q_1^{(1)}(t)\sqrt{\delta Q_2^{(0)}(t)}dB_{21}(t) \\
&\quad + \lambda_t dt + p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)dt + (1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right)dt + \delta Q_2^{(0)}(t)dt, \\
d\left(Q_2^{(1)}(t)\right)^2 &= 2Q_2^{(1)}(t)\left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \delta Q_2^{(1)}(t)\right)dt \\
&\quad + 2Q_2^{(1)}(t)\sqrt{p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_{12}(t) - 2Q_2^{(1)}(t)\sqrt{\delta Q_2^{(0)}(t)}dB_{21}(t) \\
&\quad + p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)dt + \delta Q_2^{(0)}(t)dt,
\end{aligned}$$

$$
\begin{aligned}
d\left(Q_1^{(1)}(t)Q_2^{(1)}(t)\right) &= dQ_1^{(1)}(t)\cdot Q_2^{(1)}(t) + dQ_2^{(1)}(t)\cdot Q_1^{(1)}(t) + dQ_1^{(1)}(t)\cdot dQ_2^{(1)}(t) \\
&= Q_2^{(1)}(t)\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ + \delta Q_2^{(1)}(t)\right)dt \\
&+ Q_2^{(1)}(t)\sqrt{\lambda_t}dB_1^a(t) - Q_2^{(1)}(t)\sqrt{p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_2^d(t) \\
&- Q_2^{(1)}(t)\sqrt{(1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_{12}(t) + Q_2^{(1)}(t)\sqrt{\delta Q_2^{(0)}(t)}dB_{21}(t) \\
&+ Q_1^{(1)}(t)\left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \delta Q_2^{(1)}(t)\right)dt \\
&+ Q_1^{(1)}(t)\sqrt{(1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right)}dB_{12}(t) - Q_1^{(1)}(t)\sqrt{\delta Q_2^{(0)}(t)}dB_{21}(t) \\
&- p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)dt - \delta Q_2^{(0)}(t)dt.
\end{aligned}
$$

Taking the expectation of these causes all the terms containing the BM to disappear, thus:

$$
\begin{aligned}
&\frac{d}{dt}\mathrm{E}\left[\left(Q_1^{(1)}(t)\right)^2\right] = 2\mathrm{E}\left[Q_1^{(1)}(t)\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ + \delta Q_2^{(1)}(t)\right)\right] \\
&+ \lambda_t + p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + (1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t) \\
&\frac{d}{dt}\mathrm{E}\left[\left(Q_2^{(1)}(t)\right)^2\right] = 2\mathrm{E}\left[Q_2^{(1)}(t)\left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \delta Q_2^{(1)}(t)\right)\right] \\
&+ p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t) \\
&\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] = \mathrm{E}\left[Q_2^{(1)}(t)\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ + \delta Q_2^{(1)}(t)\right)\right] \\
&+ \mathrm{E}\left[Q_1^{(1)}(t)\left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \delta Q_2^{(1)}(t)\right)\right] \\
&- p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) - \delta Q_2^{(0)}(t)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{d}{dt}\mathrm{Var}\left[Q_1^{(1)}(t)\right] = \frac{d}{dt}\mathrm{E}\left[\left(Q_1^{(1)}(t)\right)^2\right] - 2\mathrm{E}\left[Q_1^{(1)}(t)\right]\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)\right] \\
&= 2\mathrm{E}\left[Q_1^{(1)}(t)\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ + \delta Q_2^{(1)}(t)\right)\right] \\
&+ \lambda_t + p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + (1-p)\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t) \\
&- 2\mathrm{E}\left[Q_1^{(1)}(t)\right]\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right] + \delta\mathrm{E}\left[Q_2^{(1)}(t)\right]\right), \\
&\frac{d}{dt}\mathrm{Var}\left[Q_2^{(1)}(t)\right] = 2\mathrm{E}\left[Q_2^{(1)}(t)\left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}Q_1^{(1)}(t)^+ - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- - \delta Q_2^{(1)}(t)\right)\right] \\
&+ p\mu\left(Q_1^{(0)}(t)\wedge s_t\right) + \delta Q_2^{(0)}(t) \\
&- 2\mathrm{E}\left[Q_2^{(1)}(t)\right]\left(p\mu 1_{\left\{Q_1^{(0)}(t)< s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right] - p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right] - \delta\mathrm{E}\left[Q_2^{(1)}(t)\right]\right)
\end{aligned}
$$

$$\frac{d}{dt}\mathrm{Cov}\left[Q_1^{(1)}(t),Q_2^{(1)}(t)\right]=\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)Q_2^{(1)}(t)\right]-\frac{d}{dt}\mathrm{E}\left[Q_1^{(1)}(t)\right]\cdot\mathrm{E}\left[Q_2^{(1)}(t)\right]-\mathrm{E}\left[Q_1^{(1)}(t)\right]\cdot\frac{d}{dt}\mathrm{E}\left[Q_2^{(1)}(t)\right]$$
$$=\mathrm{E}\left[Q_2^{(1)}(t)\left(\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- -\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}Q_1^{(1)}(t)^+ +\delta Q_2^{(1)}(t)\right)\right]$$
$$+\mathrm{E}\left[Q_1^{(1)}(t)\left(p\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}Q_1^{(1)}(t)^+ -p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}Q_1^{(1)}(t)^- -\delta Q_2^{(1)}(t)\right)\right]$$
$$-p\mu\left(Q_1^{(0)}(t)\wedge s_t\right)-\delta Q_2^{(0)}(t)$$
$$-\mathrm{E}\left[Q_2^{(1)}(t)\right]\left[\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right]-\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right]+\delta\mathrm{E}\left[Q_2^{(1)}(t)\right]\right]$$
$$-\mathrm{E}\left[Q_1^{(1)}(t)\right]\left[p\mu 1_{\left\{Q_1^{(0)}(t)<s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^+\right]-p\mu 1_{\left\{Q_1^{(0)}(t)\le s_t\right\}}\mathrm{E}\left[Q_1^{(1)}(t)^-\right]-\delta\mathrm{E}\left[Q_2^{(1)}(t)\right]\right].$$

Rearranging some of the terms yields expression (B.7).

□

B.4 The Virtual Waiting Time Process: Proofs

Proof of Theorem 26 in Section B.2.

Proof. Introduce the processes A_i^η and D_i^η which represent the arrival and departure processes of station $i\in\{1,2\}$, respectively.

$$\begin{aligned}
A_1^\eta(t)&=A_1^a\left(\int_0^t\eta\lambda_u du\right)+A_{21}\left(\int_0^t\delta Q_2^\eta(u)du\right),\\
A_2^\eta(t)&=A_{12}\left(\int_0^t p\mu\left(Q_1^\eta(u)\wedge\eta s_u\right)du\right),\\
D_1^\eta(t)&=A_2^d\left(\int_0^t p\mu\left(Q_1^\eta(u)\wedge\eta s_u\right)du\right)+A_{12}\left(\int_0^t(1-p)\mu\left(Q_1^\eta(u)\wedge\eta s_u\right)du\right),\\
D_2^\eta(t)&=A_{21}\left(\int_0^t\delta Q_2^\eta(u)du\right).
\end{aligned}\tag{B.10}$$

Naturally, this is based on the following relation:

$$Q^\eta(t)=Q^\eta(0)+A^\eta(t)-D^\eta(t),\ t\ge 0.$$

Let $Z^\eta(t)$ be the process:

$$Z^\eta(t)=\inf\{u\ge 0: D_1^\eta(u)\ge Q_1^\eta(0)+A_1^\eta(t)-(\eta s_t-1)\}.$$

and let $W^\eta(t)$ be the virtual waiting time at t, i.e., $W^\eta(t)=[Z^\eta(t)-t]^+$. Let for $t\ge 0$,

$$X_i^\eta(t)=\frac{1}{\eta}D_i^\eta,\quad Y_i^\eta(t)=\frac{1}{\eta}A_i^\eta,\quad i=1,2.$$

Consequently,

$$Z^\eta(t)=\inf\{u\ge 0: X_1^\eta(u)\ge Y_1^\eta(t)-\frac{1}{\eta}(\eta s_t-1)+\frac{1}{\eta}Q_1^\eta(0)\}.$$

Since (B.4) holds in fact the following holds

$$\left(\sqrt{\eta}\left(\frac{Q^{\eta}(t)}{\eta}-Q^{(0)}(t)\right),\ \sqrt{\eta}\left(X^{\eta}(t)-X^{(0)}(t)\right),\ \sqrt{\eta}\left(Y^{\eta}(t)-Y^{(0)}(t)\right)\right)\xrightarrow{d}\left(Q^{(1)}(t),U(t),V(t)\right)$$

where

$$\begin{aligned}
Q^{(0)}(t)&=Q^{(0)}(0)+Y^{(0)}(t)-X^{(0)}(t),\\
X_1^{(0)}(t)&=\int_0^t\mu\left(Q_1^{(0)}(u)\wedge s_u\right)du,\\
X_2^{(0)}(t)&=\int_0^t\delta Q_2^{(0)}(u)du,\\
Y_1^{(0)}(t)&=\int_0^t\left(\lambda_u+\delta Q_2^{(0)}(u)\right)du,\\
Y_2^{(0)}(t)&=\int_0^t p\mu\left(Q_1^{(0)}(u)\wedge s_u\right)du,\\
U_1(t)&=\int_0^t\left(\mu 1_{\left\{Q_1^{(0)}(u)<s_u\right\}}Q_1^{(1)}(u)^+-\mu 1_{\left\{Q_1^{(0)}(u)\le s_u\right\}}Q_1^{(1)}(u)^-\right)du\\
&\quad+B_2^d\left(\int_0^t p\mu\left(Q_1^{(0)}(u)\wedge s_u\right)du\right)+B_{12}\left(\int_0^t(1-p)\mu\left(Q_1^{(0)}(u)\wedge s_u\right)du\right),\\
U_2(t)&=\int_0^t\left(\delta Q_2^{(1)}(u)\right)du+B_{21}\left(\int_0^t\delta Q_2^{(0)}(u)du\right),\\
V_1(t)&=\int_0^t\left(\delta Q_2^{(1)}(u)\right)du+B_1^a\left(\int_0^t\lambda_u du\right)+B_{21}\left(\int_0^t\delta Q_2^{(0)}(u)du\right),\\
V_2(t)&=\int_0^t\left(p\mu 1_{\left\{Q_1^{(0)}(u)<s_u\right\}}Q_1^{(1)}(u)^+-p\mu 1_{\left\{Q_1^{(0)}(u)\le s_u\right\}}Q_1^{(1)}(u)^-\right)du\\
&\quad+B_{12}\left(\int_0^t p\mu\left(Q_1^{(0)}(u)\wedge s_u\right)du\right).
\end{aligned}$$

As a consequence,

$$Q^{(1)}(t)=Q^{(1)}(0)+V(t)-U(t),\ t\ge 0.$$

Defining a first passage time:

$$Z(t)=\inf\{u\ge 0: X_1^{(0)}(u)\ge Y_1^{(0)}(t)-s_t+Q_1^{(0)}(0)\}$$

then by the corollary of Puhalskii [65] we obtain:

$$\left(\sqrt{\eta}\left(\frac{Q^{\eta}(t)}{\eta}-Q^{(0)}(t)\right),\ \sqrt{\eta}\left(Z^{\eta}(t)-Z(t)\right)\right)\xrightarrow{d}\left(Q^{(1)}(t),Z^{(1)}(t)\right)$$

where:

$$Z^{(1)}(t)=\frac{Q_1^{(1)}(0)+V_1(t)-U_1(Z(t))}{X_1'(Z(t))}=\frac{Q_1^{(1)}(0)+V_1(t)-U_1(Z(t))}{\mu\left(Q_1^{(0)}(Z(t))\wedge s_{Z(t)}\right)}.$$

and by continuance mapping:

$$\sqrt{\eta}(W^{\eta}(t)) = \sqrt{\eta}([Z^{\eta}(t) - t]^{+}) \xrightarrow{d} \left[\frac{Q_1^{(1)}(0) + V_1(t) - U_1(Z(t))}{\mu \left(Q_1^{(0)}(Z(t)) \wedge s_{Z(t)} \right)} \right]^{+}.$$

□

Note that when we use MOL staffing, the fluid solution is the same as that of the Infinite Server system. In such a system, $Z(t) = t$ for all t. In this case, approximating $E[W]$ by the fluid solution will be wrong, since $W^{(0)}(t) = 0$, while in reality it is not. That is because under the QED regime, average waiting is on the order of $\sqrt{(}s)$, and thus not captured by the fluid processes. However, the diffusion process is very interesting in this case.

$$\sqrt{\eta}(W^{\eta}(t)) \xrightarrow{d} \left[\frac{Q_1^{(1)}(t)}{\mu \left(Q_1^{(0)}(t) \wedge s_t \right)} \right]^{+}.$$

Note that we also need to adjust $Q^{(1)}(t)$ to the QED environment. This will be done in future research.

C Appendices of Part II

C.1 Steady State Calculations of MU Model

We will rewrite Expression 11.1 of π_0 . Let l be the number of occupied beds (with patients or in cleaning) and m be the number of patients, where i of them are in the needy state ($0 \leq i \leq m \leq l \leq n$). Thus, $l = i + j + k$ and $m = i + j$. Then:

$$
\begin{aligned}
\pi_0^{-1} &= \sum_{l=0}^{n} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{\nu(i)} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&= \sum_{l=0}^{s} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&+ \sum_{l=s+1}^{n} \left(\sum_{m=0}^{s} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right. \\
&\quad + \sum_{m=s+1}^{l} \left(\sum_{i=0}^{s} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right. \\
&\quad\quad \left. \left. + \sum_{i=s+1}^{m} \frac{1}{s!s^{i-s}} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right) \right) \\
&= \sum_{l=0}^{s} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&+ \sum_{l=s+1}^{n} \left(\sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right. \\
&\quad - \sum_{m=s+1}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&\quad + \sum_{m=s+1}^{l} \left(\sum_{i=0}^{s} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right. \\
&\quad\quad \left. \left. + \sum_{i=s+1}^{m} \frac{1}{s!s^{i-s}} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right) \right) \\
&= \sum_{l=0}^{s} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&+ \sum_{l=s+1}^{n} \left(\sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right. \\
&\quad - \sum_{m=s+1}^{l} \sum_{i=s+1}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&\quad \left. + \sum_{m=s+1}^{l} \sum_{i=s+1}^{m} \frac{1}{s!s^{i-s}} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \right) \\
&= \sum_{l=0}^{s} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&+ \sum_{l=s+1}^{n} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&+ \sum_{l=s+1}^{n} \sum_{m=s+1}^{l} \sum_{i=s+1}^{m} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&= \sum_{l=0}^{n} \sum_{m=0}^{l} \sum_{i=0}^{m} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m} \\
&+ \sum_{l=s+1}^{n} \sum_{m=s+1}^{l} \sum_{i=s+1}^{m} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!}\right) \left(\frac{\lambda}{(1-p)\mu}\right)^{i} \frac{1}{(m-i)!} \left(\frac{p\lambda}{(1-p)\delta}\right)^{m-i} \frac{1}{(l-m)!} \left(\frac{\lambda}{\gamma}\right)^{l-m}
\end{aligned}
$$

By using the multinomial theorem for the first sum yields:

$$\pi_0^{-1} = \sum_{l=0}^{n} \frac{1}{l!} \left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} \right)^l$$
$$+ \sum_{l=s+1}^{n} \sum_{m=s+1}^{l} \sum_{i=s+1}^{m} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!} \right) \frac{1}{(m-i)!(l-m)!} \left(\frac{\lambda}{(1-p)\mu} \right)^i \cdot$$
$$\cdot \left(\frac{p\lambda}{(1-p)\delta} \right)^{m-i} \left(\frac{\lambda}{\gamma} \right)^{l-m}$$

C.2 Four Auxiliary Lemmas

In this section we will prove four lemmas that will help us in the proofs of our approximations.

C.2.1 Proof of Lemma 1

Lemma 1. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define ζ_1 as the expression*

$$\zeta_1 = \frac{e^{-R_N}}{s!} (R_N)^s \frac{1}{1-\rho} \sum_{l=0}^{n-s-1} \frac{1}{l!} (R_D + R_C)^l e^{-(R_D+R_C)}.$$

Then

$$\lim_{\lambda\to\infty} \zeta_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta}.$$

Proof. By using Stirling's formula $\left(s! \approx \sqrt{2\pi s}\left(\frac{s}{e}\right)^s\right)$, and assumption QED (ii), one obtains for ζ_1:

$$\begin{aligned}
\zeta_1 &\approx \frac{e^{s-\frac{\lambda}{(1-p)\mu}}}{\sqrt{2\pi s}} \left(\frac{\lambda}{(1-p)s\mu} \right)^s \frac{\sqrt{s}}{\beta} \sum_{l=0}^{n-s-1} \frac{1}{l!} \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} \right)^l e^{-\left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \\
&= \frac{e^{s-\frac{\lambda}{(1-p)\mu}}}{\sqrt{2\pi}\beta} \left(\frac{\lambda}{(1-p)s\mu} \right)^s \sum_{l=0}^{n-s-1} \frac{1}{l!} \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} \right)^l e^{-\left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \\
&= \frac{e^{s(1-\rho)}}{\sqrt{2\pi}\beta} \rho^s P(X_\lambda \le n-s-1)
\end{aligned}$$

where $\rho = \frac{\lambda}{(1-p)s\mu}$, and X_λ is a random variable with the Poisson distribution with parameter $R_D + R_C$ (where $R_D = \frac{p\lambda}{(1-p)\delta}$, $R_C = \frac{\lambda}{\gamma}$). When $\lambda \to \infty$, $R_D + R_C \to \infty$ too, since p,δ, and γ are fixed. Note that

$$P(X_\lambda \le n-s-1) = P\left(\frac{X_\lambda - R_D - R_C}{\sqrt{R_D + R_C}} \le \frac{n-s-1-R_D-R_C}{\sqrt{R_D+R_C}} \right)$$

Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\left(\frac{X_\lambda - R_D - R_C}{\sqrt{R_D + R_C}} \right) \Rightarrow N(0,1)$$

and due to assumption QED (i) of the lemma we get [5]

$$P(X_\lambda \leq n-s-1) \to P(N(0,1) \leq \eta) = \Phi(\eta), \quad as\lambda \to \infty \tag{C.1}$$

where $N(0,1)$ is a standard normal random variable with distribution function $\Phi(\cdot)$. It follows thus that

$$\zeta_1 \approx \frac{e^{s(1-\rho)}}{\sqrt{2\pi}\beta}\rho^s\Phi(\eta) = \frac{e^{s(1-\rho+\ln\rho)}}{\sqrt{2\pi}\beta}\Phi(\eta).$$

Making use of the expansion

$$\ln\rho = \ln(1-(1-\rho)) = -(1-\rho) - \frac{(1-\rho)^2}{2} + o(1-\rho)^2, \quad (\rho \to 1)$$

one obtains

$$\zeta_1 \approx \frac{e^{s(1-\rho-(1-\rho)-\frac{(1-\rho)^2}{2})}}{\sqrt{2\pi}\beta}\Phi(\eta) = \frac{e^{-\frac{s(1-\rho)^2}{2}}}{\sqrt{2\pi}\beta}\Phi(\eta)$$

by assumption QED (ii) $s(1-\rho)^2 \to \beta^2$, when $\lambda \to \infty$. This implies

$$\lim_{\lambda\to\infty}\zeta_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta}$$

where $\phi(\cdot)$ is the standard normal density function, and $\Phi(\cdot)$ is the standard normal distribution function. This proves Lemma 1. □

C.2.2 Proof of Lemma 2

Lemma 2. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define ζ_2 as the expression*

$$\zeta_2 = \frac{e^{-(R_N+R_D+R_C)}}{s!}(R_N)^s\frac{\rho^{n-s}}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{R_D}{\rho}+\frac{R_C}{\rho}\right)^l. \tag{C.2}$$

Then

$$\lim_{\lambda\to\infty}\zeta_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1).$$

[5] Here we use the following theorem (from [7]):

Theorem 27. *Let $\varsigma_n \Rightarrow \varsigma$ and F_ς - the distribution function of ς is everywhere continuous. Let also $x_n \to x_\infty$ as $n \to \infty$, where $\{x_n\}$ is a sequence of scalars. Here $x_\infty \in [-\infty,\infty]$. Then $F_{\varsigma_n}(x_n) \to F_\varsigma(x_\infty)$.*

Proof. Again according to Stirling's formula, and assumption QED (ii), one obtains for ζ_2:

$$\begin{aligned}\zeta_2 &\approx \frac{e^{s-\frac{\lambda}{(1-p)\mu}-\frac{p\lambda}{(1-p)\delta}-\frac{\lambda}{\gamma}}}{\sqrt{2\pi s}}\left(\frac{\lambda}{(1-p)s\mu}\right)^s\frac{\rho^{n-s}}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)^l\\ &= \frac{e^{s(1-\rho)-\frac{p\lambda}{(1-p)\delta}-\frac{\lambda}{\gamma}}}{\sqrt{2\pi s}}\frac{\sqrt{s}\rho^n}{\beta}e^{\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)^l e^{-\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)}\\ &= \frac{e^{s(1-\rho)+\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)\left(\frac{1-\rho}{\rho}\right)}}{\sqrt{2\pi}\beta}\rho^n P(Y_\lambda \le n-s-1)\end{aligned}$$

where $\rho=\frac{\lambda}{(1-p)s\mu}$, and Y_λ is a random variable with the Poisson distribution with parameter $\frac{R_D+R_C}{\rho}$ (where $R_D=\frac{p\lambda}{(1-p)\delta}$, $R_C=\frac{\lambda}{\gamma}$). Note that

$$P(Y_\lambda \le n-s-1)=P\left(\frac{Y_\lambda-\frac{R_D+R_C}{\rho}}{\sqrt{\frac{R_D+R_C}{\rho}}}\le\frac{n-s-1-\frac{R_D+R_C}{\rho}}{\sqrt{\frac{R_D+R_C}{\rho}}}\right)$$

Now we need to find the limit for the following fraction

$$\frac{n-s-\frac{R_D+R_C}{\rho}}{\sqrt{\frac{R_D+R_C}{\rho}}}$$

as $\lambda\to\infty$ using assumption QED (i).

$$\begin{aligned}&\lim_{\lambda\to\infty}\frac{n-s-\frac{R_D+R_C}{\rho}}{\sqrt{\frac{R_D+R_C}{\rho}}}=\lim_{\lambda\to\infty}\frac{\eta\sqrt{R_D+R_C}+R_D+R_C-\frac{R_D+R_C}{\rho}}{\sqrt{\frac{R_D+R_C}{\rho}}}\\ &=\lim_{\lambda\to\infty}\eta\sqrt{\rho}+\frac{\sqrt{R_D+R_C}(\rho-1)}{\sqrt{\rho}}=\eta-\lim_{\lambda\to\infty}\sqrt{\frac{s\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}(1-\rho)\\ &=\eta-\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}\beta.\end{aligned}\tag{C.3}$$

Denote

$$\eta_1=\eta-\beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}$$

Thus, when $\lambda\to\infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\left(\frac{Y_\lambda-\frac{R_D-R_C}{\rho}}{\sqrt{\frac{R_D+R_C}{\rho}}}\right)\Rightarrow N(0,1)$$

and

$$P(Y_\lambda\le n-s-1)\to P(N(0,1)\le\eta_1)=\Phi(\eta_1),\text{ as }\lambda\to\infty$$

where $N(0,1)$ is a standard normal random variable with distribution function Φ. It follows thus that

$$\zeta_2 \approx \frac{e^{s(1-\rho)+\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)\left(\frac{1-\rho}{\rho}\right)}}{\sqrt{2\pi}\beta}\rho^n\Phi(\eta_1) = \frac{e^{s(1-\rho)+\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)\left(\frac{1-\rho}{\rho}\right)+n\ln\rho}}{\sqrt{2\pi}\beta}\Phi(\eta_1).$$

Making use of the expansion

$$\ln\rho = \ln(1-(1-\rho)) = -(1-\rho) - \frac{(1-\rho)^2}{2} + o(1-\rho)^2, \quad (\rho \to 1)$$

and using our assumptions that as $\lambda \to \infty$: $\rho \to 1$, $s \approx \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}}$, and $n - s \approx \eta\sqrt{R_D+R_C} + R_D + R_C$, one obtains

$$\begin{aligned}
&s(1-\rho) + \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)\left(\frac{1-\rho}{\rho}\right) + n\ln\rho \\
&= s(1-\rho) + \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)\left(\frac{1-\rho}{\rho}\right) - n\left(1-\rho+\frac{(1-\rho)^2}{2}\right) \\
&= -\left(n - s - \frac{p\lambda}{(1-p)\delta\rho} - \frac{\lambda}{\gamma\rho}\right)(1-\rho) - \frac{n(1-\rho)^2}{2} \\
&\approx -\left(\eta\sqrt{R_D+R_C} + R_D + R_C - \frac{R_D+R_C}{\rho}\right)(1-\rho) - \frac{n(1-\rho)^2}{2} \\
&= \left(\frac{R_D+R_C}{\rho} - \frac{n}{2}\right)(1-\rho)^2 - \eta\sqrt{R_D+R_C}(1-\rho) \\
&\approx \left(\frac{R_D+R_C}{\rho} - \frac{\eta\sqrt{R_D+R_C} + R_D + R_C + \frac{\lambda}{(1-p)\mu} + \beta\sqrt{\frac{\lambda}{(1-p)\mu}}}{2}\right)(1-\rho)^2 - \eta\sqrt{R_D+R_C}(1-\rho) \\
&= \left(\frac{R_D+R_C}{\rho} - \frac{R_D+R_C+\frac{\lambda}{(1-p)\mu}}{2}\right)(1-\rho)^2 - \frac{1}{2}\beta\sqrt{\frac{\lambda}{(1-p)\mu}}(1-\rho)^2 \\
&- \frac{\eta\sqrt{R_D+R_C}}{2}(1-\rho)^2 - \eta\sqrt{R_D+R_C}(1-\rho) \\
&= \left(\frac{R_D+R_C}{\rho} - \frac{R_D+R_C+\frac{\lambda}{(1-p)\mu}}{2}\right)\frac{\beta^2(1-p)\mu}{\lambda} - \frac{1}{2}\beta\sqrt{\frac{\lambda}{(1-p)\mu}}\frac{\beta^2(1-p)\mu}{\lambda} \\
&- \frac{\eta\sqrt{R_D+R_C}}{2}\frac{\beta^2(1-p)\mu}{\lambda} - \eta\sqrt{R_D+R_C}\beta\sqrt{\frac{(1-p)\mu}{\lambda}} \\
&\approx \left(\frac{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}}{\rho} - \frac{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}+\frac{1}{(1-p)\mu}}{2}\right)(\beta^2(1-p)\mu) - \eta\sqrt{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}}\beta\sqrt{(1-p)\mu} \\
&\approx \frac{1}{2}\beta^2\left(\left(\frac{p}{(1-p)\delta}+\frac{1}{\gamma}\right)(1-p)\mu - 1\right) - \eta\beta\sqrt{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}}\sqrt{(1-p)\mu}
\end{aligned}$$

$$= -\frac{1}{2}(\eta^2+\beta^2) + \frac{1}{2}\left(\eta^2 - 2\left(\eta\beta\sqrt{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}}\sqrt{(1-p)\mu}\right)\right.$$
$$\left. + \left(\beta\sqrt{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}}\sqrt{(1-p)\mu}\right)^2\right)$$
$$= -\frac{1}{2}(\eta^2+\beta^2) + \frac{1}{2}\left(\eta - \beta\sqrt{\frac{p}{(1-p)\delta}+\frac{1}{\gamma}}\sqrt{(1-p)\mu}\right)^2 = -\frac{1}{2}(\eta^2+\beta^2) + \frac{1}{2}\eta_1^2.$$

Therefore,

$$\lim_{\lambda\to\infty}\zeta_2 \approx \frac{e^{s(1-\rho)+\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)\left(\frac{1-\rho}{\rho}\right)+n\ln\rho}}{\sqrt{2\pi}\beta}\Phi(\eta_1) \approx \frac{e^{-\frac{1}{2}(\eta^2+\beta^2)+\frac{1}{2}\eta_1^2}}{\sqrt{2\pi}\beta}\Phi(\eta_1)$$
$$= \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1).$$

This proves Lemma 2. □

C.2.3 Proof of Lemma 3

Lemma 3. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED or QED_0 conditions. Define ξ as the expression*

$$\xi = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n-1}} \frac{1}{i!j!k!}\left(R_N\right)^i\left(R_D\right)^j\left(R_C\right)^k e^{-(R_N+R_D+R_C)}.$$

Then

$$\lim_{\lambda\to\infty}\xi = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t).$$

Proof. We will find the asymptotic behavior of ξ by finding its lower and upper bounds. Let us consider a partition $\{s_h\}_{h=0}^{l}$ of the interval $[0,s]$.

$$s_h = s - h\tau, \quad h = 0,1,...,\ell; \quad s_{\ell+1} = 0$$

where $\tau = \left[\epsilon\sqrt{\frac{\lambda}{(1-p)\mu}}\right]$, ϵ is an arbitrary non-negative real and ℓ is a positive integer.

If λ and s tend to infinity and satisfy the *QED* assumption (13.2) part (ii), then $\ell < \frac{s}{\tau}$ for λ big enough and all the s_h belong to $[0,s]$; $h = 0,1,...,\ell$. Emphasize that the length τ of every interval $[s_{h-1}, s_h]$ depends on λ. The variable ξ is given by the formula (14.3). Let us consider a

lower estimate for ξ given by the following sum:

$$\begin{aligned}\xi \geq \xi_1 &= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^i e^{-\frac{\lambda}{(1-p)\mu}} \cdot \\ &\qquad \cdot \sum_{j=0}^{n-s_h-1} \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta}\right)^j e^{-\frac{p\lambda}{(1-p)\delta}} \sum_{k=0}^{n-s_h-j-1} \frac{1}{k!} \left(\frac{\lambda}{\gamma}\right)^k e^{-\frac{\lambda}{\gamma}} \\ &= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu}\right)^i e^{-\frac{\lambda}{(1-p)\mu}} P(Y_n \leq n - s_h - 1) \\ &= \sum_{h=0}^{\ell} P(s_{h+1} \leq X_n \leq s_h) P(Y_n \leq n - s_h - 1)\end{aligned} \tag{C.4}$$

where X_n and Y_n are independent Poisson random variables with parameters $\frac{\lambda}{(1-p)\mu}$ and $\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}$, respectively.

If $\lambda \to \infty$ then $\frac{\lambda}{(1-p)\mu} \to \infty$, since p and μ are fixed. Note that

$$P(s_{h+1} \leq X_n \leq s_h) = P\left(\frac{s_{h+1} - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}} \leq \frac{X_n - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}} \leq \frac{s_h - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}}\right).$$

Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\frac{X_n - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}} \Rightarrow N(0,1).$$

Since

$$\begin{aligned}\lim_{\lambda\to\infty} \frac{s_h - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}} &= \lim_{\lambda\to\infty} \frac{s - h\epsilon\sqrt{\frac{\lambda}{(1-p)\mu}} - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}} \\ = \lim_{\lambda\to\infty} \frac{\frac{\lambda}{(1-p)\mu} - \beta\sqrt{\frac{\lambda}{(1-p)\mu}} - h\epsilon\sqrt{\frac{\lambda}{(1-p)\mu}} - \frac{\lambda}{(1-p)\mu}}{\sqrt{\frac{\lambda}{(1-p)\mu}}} &= \beta - h\epsilon\end{aligned}$$

we obtain:

$$\begin{aligned}&P(s_{h+1} \leq X_n \leq s_h) = \Phi(\beta - h\epsilon) - \Phi(\beta - (h+1)\epsilon),\ h = 0, .., \ell - 1 \\ &P(0 \leq X_n \leq s_\ell) = \Phi(\beta - \ell\epsilon).\end{aligned} \tag{C.5}$$

Similarly, if $\lambda \to \infty$ then $\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} \to \infty$, since p,δ and γ are fixed. Note that

$$P(Y_n \leq n - s_h) = P\left(\frac{Y_n - \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} \leq \frac{n - s_h - \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}}\right).$$

Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\frac{Y_n - \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} \Rightarrow N(0,1).$$

Since

$$\begin{aligned}
&\lim_{\lambda\to\infty} \frac{n - s_h - \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} = \lim_{\lambda\to\infty} \frac{n - s - h\epsilon\sqrt{\frac{\lambda}{(1-p)\mu}} - \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} \\
&= \lim_{\lambda\to\infty} \frac{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} - h\epsilon\sqrt{\frac{\lambda}{(1-p)\mu}} - \left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} \\
&= \eta - h\epsilon\frac{\sqrt{\frac{1}{(1-p)\mu}}}{\sqrt{\frac{p}{(1-p)\delta} + \frac{1}{\gamma}}} = \eta - h\epsilon\sqrt{\frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}},
\end{aligned}$$

we obtain:

$$P(Y_n \le n - s_h) = \Phi\left(\eta - h\epsilon\sqrt{\frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}}\right), \ h = 0, .., \ell \tag{C.6}$$

It follows from (C.4), (C.5), and (C.6) that

$$\begin{aligned}
\lim_{\lambda\to\infty} \xi \ge \sum_{h=0}^{\ell-1} (\Phi(\beta - h\epsilon) - \Phi(\beta - (h+1)\epsilon))\Phi\left(\eta - h\epsilon\sqrt{\frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}}\right) \\
+ \Phi(\beta - \ell\epsilon)\Phi\left(\eta - \ell\epsilon\sqrt{\frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}}\right)
\end{aligned} \tag{C.7}$$

which is the lower Riemann-Stieltjes sum of the integral

$$\begin{aligned}
-\int_0^\infty \Phi\left(\eta + x\sqrt{\frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}}\right) d\Phi(\beta - x) \\
= \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma + (1-p)\delta)}}\right) d\Phi(t)
\end{aligned} \tag{C.8}$$

corresponding to the partition $\{\beta - h\epsilon\}_{h=0}^{\ell}$ of the semi axis $(-\infty, \beta)$.

Similarly, let us take the upper estimate for ξ as the following sum:

$$\begin{aligned}
\xi \le \xi_2 &= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i e^{-\frac{\lambda}{(1-p)\mu}} \cdot \\
&\qquad \sum_{j=0}^{n-s_{h+1}-1} \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j e^{-\frac{p\lambda}{(1-p)\delta}} \sum_{k=0}^{n-s_{h+1}-j-1} \frac{1}{k!}\left(\frac{\lambda}{\gamma}\right)^k e^{-\frac{\lambda}{\gamma}} \\
&= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i e^{-\frac{\lambda}{(1-p)\mu}} P(Y_n \le n - s_{h+1} - 1) \\
&= \sum_{h=0}^{\ell} P(s_{h+1} \le X_n \le s_h) P(Y_n \le n - s_{h+1} - 1)
\end{aligned} \tag{C.9}$$

where X_n and Y_n are the same random variable as before. Using the same calculation that were computed for the upper boundary we obtain

$$\lim_{\lambda\to\infty}\xi \leq \sum_{h=0}^{\ell-1}\left(\Phi(\beta-h\epsilon)-\Phi(\beta-(h+1)\epsilon)\right)\Phi\left(\eta-(h+1)\epsilon\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) + \Phi(\beta-\ell\epsilon) \tag{C.10}$$

which is the upper Riemann-Stieltjes sum for the integral (C.8). When $\epsilon\to 0$ the boundaries (C.7) and (C.10) lead to the following equality

$$\lim_{\lambda\to\infty}\xi = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t)$$

This proves Lemma 3. □

C.2.4 Proof of Lemma 4

Lemma 4. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Define ζ as the expression*

$$\zeta = e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}{R_D}^j{R_C}^k\sum_{i=0}^{n-s-j-k-1}\rho^i. \tag{C.11}$$

Then

$$\lim_{\lambda\to\infty}\zeta = \sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right).$$

Proof. First, we will rewrite Equation (C.11):

$$\begin{aligned}\zeta &= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}{R_D}^j{R_C}^k\sum_{i=0}^{n-s-j-k-1}\rho^i\\ &= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}{R_D}^j{R_C}^k\frac{1-\rho^{n-s-j-k}}{1-\rho}.\end{aligned}$$

When $\beta=0$ by assumption QED_0 (ii) $\rho=1$ therefore,

$$\sum_{i=0}^{n-s-j-k-1}\rho^i = n-s-j-k.$$

When $\beta\to 0$, $\rho\to 1$ but still $\rho\neq 1$, the expression $\frac{1-\rho^{n-s-j-k}}{1-\rho}$ can be approximated by

$$\frac{1-\rho^i}{1-\rho}\approx i$$

thus

$$\lim_{\rho\to 1} \sum_{i=0}^{n-s-j-k-1} \rho^i = n-s-j-k,$$

which is the same phrase as when $\rho = 1$. Thus,

$$\begin{aligned}
\zeta &= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s \sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}{R_D}^j{R_C}^k(n-s-j-k)\\
&= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\left((n-s)\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{k!j!}{R_C}^k{R_D}^j-\right.\\
&\left.-\sum_{k=0}^{n-s-1}\frac{k}{k!}{R_C}^k\sum_{j=0}^{n-s-k-1}\frac{1}{j!}{R_D}^j-\sum_{k=0}^{n-s-1}\frac{1}{k!}{R_C}^k\sum_{j=0}^{n-s-k-1}\frac{j}{j!}{R_D}^j\right)\\
&= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\left((n-s)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l-\right.\\
&\left.-R_C\sum_{k=0}^{n-s-2}\frac{1}{k!}{R_C}^k\sum_{j=0}^{n-s-k-2}\frac{1}{j!}{R_D}^j-R_D\sum_{k=0}^{n-s-2}\frac{1}{k!}{R_C}^k\sum_{j=0}^{n-s-k-2}\frac{1}{j!}{R_D}^j\right)\\
&= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\left((n-s)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l-\right.\\
&\left.-R_C\sum_{l=0}^{n-s-2}\frac{1}{l!}(R_C+R_D)^l-R_D\sum_{l=0}^{n-s-2}\frac{1}{l!}(R_C+R_D)^l\right)\\
&= e^{-(R_N+R_D+R_C)}\frac{1}{s!}{R_N}^s\left((n-s-R_C-R_D)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l+\frac{(R_C+R_D)^{n-s}}{(n-s-1)!}\right)\\
&\approx e^{s-R_N}\frac{1}{\sqrt{2\pi s}}\rho^s\left((n-s-R_C-R_D)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l e^{-(R_D+R_C)}\right.\\
&\left.+\frac{(R_C+R_D)^{n-s}e^{-(R_D+R_C)}}{(n-s-1)!}\right)\\
&= \frac{1}{\sqrt{2\pi s}}\left((n-s-R_C-R_D)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l e^{-(R_D+R_C)}+\frac{(R_C+R_D)^{n-s}e^{-(R_D+R_C)}}{(n-s-1)!}\right).
\end{aligned}$$

As seen in Equation (C.1)

$$(n-s-R_C-R_D)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l e^{-(R_D+R_C)} \approx \eta\sqrt{R_C+R_D}\Phi(\eta). \tag{C.12}$$

By using Stirling's formula:

$$\begin{aligned}
&\frac{(R_C+R_D)^{n-s}e^{-(R_D+R_C)}}{(n-s-1)!} = \frac{(n-s)(R_C+R_D)^{n-s}e^{-(R_D+R_C)}}{(n-s)!} \\
&\approx \frac{(n-s)e^{n-s-(R_D+R_C)}}{\sqrt{2\pi(n-s)}}\left(\frac{R_C+R_D}{n-s}\right)^{n-s} = \frac{(n-s)e^{n-s-(R_D+R_C)+(n-s)\ln\left(\frac{R_C+R_D}{n-s}\right)}}{\sqrt{2\pi(n-s)}} \\
&= \sqrt{\frac{n-s}{2\pi}}e^{(n-s)\left(1-\frac{R_D+R_C}{n-s}+\ln\left(\frac{R_C+R_D}{n-s}\right)\right)}.
\end{aligned}$$

By assuming QED_0 (i) when $\lambda \to \infty$

$$\begin{aligned}
&(n-s)\left(1-\frac{R_D+R_C}{n-s}+\ln\left(\frac{R_C+R_D}{n-s}\right)\right) \\
&= (n-s)\left(1-\frac{R_D+R_C}{n-s}-\left(1-\frac{R_C+R_D}{n-s}\right)-\frac{1}{2}\left(1-\frac{R_C+R_D}{n-s}\right)^2\right) \\
&= -\frac{n-s}{2}\left(1-\frac{R_C+R_D}{n-s}\right)^2 = -\frac{1}{2}\frac{(n-s-R_C-R_D)^2}{n-s} \\
&\approx -\frac{1}{2}\frac{\left(\eta\sqrt{R_C+R_D}\right)^2}{\eta\sqrt{R_C+R_D}+R_C+R_D} \approx -\frac{1}{2}\frac{\left(\eta\sqrt{R_C+R_D}\right)^2}{\eta\sqrt{R_C+R_D}+R_C+R_D} \approx -\frac{\eta^2}{2}.
\end{aligned} \tag{C.13}$$

Therefore, by assumption QED_0 (i)

$$\begin{aligned}
&\sqrt{\frac{n-s}{2\pi}}e^{(n-s)\left(1-\frac{R_D+R_C}{n-s}+\ln\left(\frac{R_C+R_D}{n-s}\right)\right)} \approx \sqrt{\frac{\eta\sqrt{R_C+R_D}+R_C+R_D}{2\pi}}e^{-\frac{\eta^2}{2}} \\
&= \sqrt{\eta\sqrt{R_C+R_D}+R_C+R_D}\phi(\eta) \approx \sqrt{R_C+R_D}\phi(\eta).
\end{aligned}$$

Combining the above approximations and the assumption that $\beta = 0$ and therefore $s = R_N = \frac{\lambda}{(1-p)\mu}$ yields

$$\begin{aligned}
\zeta &\approx \frac{1}{\sqrt{2\pi s}}\left((n-s-R_C-R_D)\sum_{l=0}^{n-s-1}\frac{1}{l!}(R_C+R_D)^l e^{-(R_D+R_C)} + \frac{(R_C+R_D)^{n-s}e^{-(R_D+R_C)}}{(n-s-1)!}\right) \\
&\approx \frac{1}{\sqrt{2\pi s}}\left(\eta\sqrt{R_C+R_D}\Phi(\eta)+\sqrt{R_C+R_D}\phi(\eta)\right) \\
&= \frac{\sqrt{R_C+R_D}}{\sqrt{2\pi s}}\left(\eta\Phi(\eta)+\phi(\eta)\right) = \frac{\sqrt{R_C+R_D}}{\sqrt{2\pi R_N}}\left(\eta\Phi(\eta)+\phi(\eta)\right) \\
&= \sqrt{\frac{R_C+R_D}{R_N}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right) = \sqrt{\frac{(1-p)\mu}{\gamma}+\frac{p\mu}{\delta}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right).
\end{aligned}$$

This proves Lemma 4. □

C.3 Proof of Approximation of the Expected Waiting Time

In this appendix we will prove the approximation for the expected waiting time, stated in Section 14.2. The accurate measure was defined in Section 12.2, by Formula (12.5).

Theorem 10. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the $QED_{\neq 0}$ conditions. Then*

$$\lim_{\lambda\to\infty}\sqrt{s}E[W]=\frac{\frac{\phi(\beta)\Phi(\eta)}{\beta}\frac{1}{\beta}+\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\left(\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}\beta-\eta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}-\frac{1}{\beta}\right)}{\mu\left(\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\right)}$$

where $\eta_1=\eta-\beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}$.

Proof. It follows from (12.5) that the expectation of the waiting time is given by

$$\begin{aligned}
E[W]&=\int_0^\infty p_n(s;t)dt=\frac{1}{\mu s}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m)(i-s+1)\\
&=\frac{1}{\mu s}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m)(i-s)+\frac{1}{\mu s}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m)\\
&=\frac{1}{\mu s}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m)(i-s)+\frac{1}{\mu s}P(W>0)\\
&=C+D
\end{aligned}\tag{C.14}$$

where D is given by

$$D=\frac{1}{\mu s}P(W>0)=\frac{1}{\mu s}\frac{B}{A+B}$$

(A and B where defined in Equation (14.1) and (14.2) respectively) and C is given by,

$$\begin{aligned}
C&=\frac{1}{\mu s}\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\pi_{n-1}(i,m-i,l-m)(i-s)\\
&=\frac{1}{\mu s}\pi_0\sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\frac{1}{s!s^{i-s}}\left(R_N\right)^i\frac{1}{(m-i)!(l-m)!}\left(R_D\right)^{m-i}\left(R_C\right)^{l-m}(i-s)\\
&=\frac{1}{\mu s}\frac{G}{A+B}.
\end{aligned}$$

We will rewrite G in the following way:

$$
\begin{aligned}
G &= \sum_{l=s}^{n-1}\sum_{m=s}^{l}\sum_{i=s}^{m}\frac{1}{s!s^{i-s}}\left(R_N\right)^i\frac{1}{(m-i)!(l-m)!}\left(R_D\right)^{m-i}\left(R_C\right)^{l-m}(i-s)\\
&= \sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\sum_{i=s}^{n-j-k-1}\frac{1}{s!s^{i-s}}\left(R_N\right)^i\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k(i-s)\\
&= \sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\sum_{i=0}^{n-s-j-k-1}\frac{i}{s!s^{i}}\left(R_N\right)^{i+s}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\\
&= \frac{\left(R_N\right)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\sum_{i=0}^{n-s-j-k-1}i\left(\rho\right)^i.
\end{aligned}
$$

Using the formula

$$
\begin{aligned}
&\sum_{l=0}^{M}l\rho^l=\rho\left(\sum_{l=0}^{M}\rho^l\right)'=\rho\left(\frac{1-\rho^{M+1}}{1-\rho}\right)'=\rho\frac{(-(M+1)\rho^M)(1-\rho)-(1-\rho^{M+1})(-1)}{(1-\rho)^2}=\\
&=(M+1)\frac{\rho^{M+1}}{\rho-1}+\frac{1-\rho^{M+1}}{(1-\rho)^2}\rho=...=M\frac{\rho^{M+1}}{\rho-1}+\frac{1-\rho^M}{(1-\rho)^2}\rho
\end{aligned}
\tag{C.15}
$$

one can rewrite G as a sum: $G=G_1+G_2$; where

$$
G_1=\frac{\left(R_N\right)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\left((n-s-j-k-1)\frac{\rho^{n-s-j-k}}{\rho-1}\right)
$$

and

$$
G_2=\frac{\left(R_N\right)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\left(\frac{1-\rho^{n-s-j-k-1}}{(1-\rho)^2}\rho\right).
$$

Therefore,

$$
\begin{aligned}
G_1 &= \frac{\left(R_N\right)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\left((n-s-j-k-1)\frac{\rho^{n-s-j-k}}{\rho-1}\right)\\
&= \frac{\left(R_N\right)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\left((n-s-1)\frac{\rho^{n-s-j-k}}{\rho-1}\right)\\
&- \frac{\left(R_N\right)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(R_D\right)^j\left(R_C\right)^k\left((j+k)\frac{\rho^{n-s-j-k}}{\rho-1}\right)\\
&= \frac{\left(R_N\right)^s(n-s-1)}{s!}\frac{\rho^{n-s}}{\rho-1}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}\left(\frac{R_D}{\rho}\right)^j\left(\frac{R_C}{\rho}\right)^k\\
&- \frac{\left(R_N\right)^s}{s!}\frac{\rho^{n-s}}{\rho-1}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{j+k}{j!k!}\left(\frac{R_D}{\rho}\right)^j\left(\frac{R_C}{\rho}\right)^k
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{(R_N)^s (n-s-1)}{s!} \frac{\rho^{n-s}}{\rho-1} \sum_{l=0}^{n-s-1} \frac{1}{l!} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right)^l \\
&- \frac{(R_N)^s}{s!} \frac{\rho^{n-s}}{\rho-1} \sum_{l=0}^{n-s-1} \frac{l}{l!} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right)^l \\
&= \frac{(R_N)^s (n-s-1)}{s!} \frac{\rho^{n-s}}{\rho-1} \sum_{l=0}^{n-s-1} \frac{1}{l!} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right)^l \\
&- \frac{(R_N)^s}{s!} \frac{\rho^{n-s}}{\rho-1} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right) \sum_{l=0}^{n-s-2} \frac{l}{l!} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right)^l \\
&= -(n-s-1) e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \zeta_2 + \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right) e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \zeta_2 \\
&+ \frac{(R_N)^s}{s!} \frac{\rho^{n-s}}{\rho-1} \frac{1}{(n-s-1)!} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right)^{n-s} \\
&= -(n-s-1) e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \zeta_2 + \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right) e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \zeta_2 \\
&+ \frac{(R_N)^s}{s!} \frac{1}{\rho-1} \frac{n-s}{(n-s)!} (R_D + R_C)^{n-s}
\end{aligned}
$$

where ζ_2 was defined in (14.5); and

$$
\begin{aligned}
G_2 &= \frac{(R_N)^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k \left(\frac{1-\rho^{n-s-j-k-1}}{(1-\rho)^2}\rho\right) \\
&= \frac{\rho}{(1-\rho)^2} \frac{(R_N)^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k \\
&- \frac{\rho^{n-s}}{(1-\rho)^2} \frac{(R_N)^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} \left(\frac{R_D}{\rho}\right)^j \left(\frac{R_C}{\rho}\right)^k \\
&= \frac{\rho}{(1-\rho)^2} \frac{(R_N)^s}{s!} \sum_{l=0}^{n-s-1} \frac{1}{l!} (R_D + R_C)^l \\
&- \frac{\rho^{n-s}}{(1-\rho)^2} \frac{(R_N)^s}{s!} \sum_{l=0}^{n-s-1} \frac{1}{l!} \left(\frac{R_D}{\rho} + \frac{R_C}{\rho}\right)^l \\
&= \frac{\rho}{(1-\rho)} e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \zeta_1 - \frac{1}{(1-\rho)} e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} \zeta_2 \\
&= \frac{1}{(1-\rho)} e^{\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)} (\rho\zeta_1 - \zeta_2)
\end{aligned}
$$

where ζ_1 was defined in (14.4). Multiplying G by $e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$ we get

$$
\begin{aligned}
&Ge^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} = \frac{1}{(1-\rho)}(\rho\zeta_1-\zeta_2)-(n-s-1)\zeta_2+\left(\frac{R_D}{\rho}+\frac{R_C}{\rho}\right)\zeta_2 \\
&+\frac{(R_N)^s}{s!}\frac{1}{\rho-1}\frac{n-s}{(n-s)!}\left(R_D+R_C\right)^{n-s}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} \\
&\approx \frac{\sqrt{s}}{\beta}\left(\left(1-\frac{\beta}{\sqrt{s}}\right)\zeta_1-\zeta_2\right)-\left(\eta\sqrt{R_C+R_D}+R_C+R_D-1\right)\zeta_2+\left(\frac{R_D}{\rho}+\frac{R_C}{\rho}\right)\zeta_2 \\
&+\frac{(R_N)^s}{s!}\frac{1}{\rho-1}\frac{n-s}{(n-s)!}\left(R_D+R_C\right)^{n-s}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} \\
&=\zeta_1\left(\frac{\sqrt{s}}{\beta}-1\right)+\zeta_2\left(-\frac{\sqrt{s}}{\beta}-\eta\sqrt{R_C+R_D}+\left(\frac{1-\rho}{\rho}\right)\left(R_C+R_D\right)-1\right) \\
&+\frac{n-s}{\rho-1}\frac{(R_N)^s}{s!}e^{-R_N}\frac{(R_D+R_C)^{n-s}}{(n-s)!}e^{-(R_D+R_C)} \\
&\approx\sqrt{s}\left(\zeta_1\frac{1}{\beta}+\zeta_2\left(\frac{R_C+R_D}{R_N}\beta-\eta\sqrt{\frac{R_C+R_D}{R_N}}-\frac{1}{\beta}\right)\right).
\end{aligned}
\tag{C.16}
$$

Due to the following approximation, we can neglect the second term:

$$
\begin{aligned}
&\frac{(R_N)^s}{s!}\frac{1}{\rho-1}\frac{n-s}{(n-s)!}\left(R_D+R_C\right)^{n-s}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} \\
&\approx \frac{(R_N)^s}{s!}\frac{1}{\rho-1}\frac{n-s}{\sqrt{2\pi(n-s)}e^{-(n-s)}(n-s)^{n-s}}\left(R_D+R_C\right)^{n-s}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} \\
&=\frac{(R_N)^s}{s!}\frac{1}{\rho-1}\sqrt{\frac{n-s}{2\pi}}\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)^{n-s}e^{n-s-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} \\
&=\frac{(R_N)^s}{s!}\frac{1}{\rho-1}\sqrt{\frac{n-s}{2\pi}}e^{n-s-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)+(n-s)\ln\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)} \\
&=\frac{(R_N)^s}{s!}\frac{1}{\rho-1}\sqrt{\frac{n-s}{2\pi}}e^{(n-s)\left(1-\left(\frac{\lambda}{(n-s)(1-p)\mu}+\frac{p\lambda}{(n-s)(1-p)\delta}+\frac{\lambda}{(n-s)\gamma}\right)+\ln\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)\right)} \\
&\approx \frac{(R_N)^s}{s!}\frac{1}{\rho-1}\sqrt{\frac{n-s}{2\pi}}e^{-R_N-\frac{\eta^2}{2}}=\frac{\sqrt{n-s}}{\rho-1}\frac{(R_N)^s}{s!}e^{-R_N}\frac{1}{\sqrt{2\pi}}e^{-\frac{\eta^2}{2}} \\
&\approx \frac{\sqrt{R_C+R_D}}{\rho-1}\frac{1}{R_N}\phi(R_N+\beta\sqrt{R_N})\phi\left(\eta\right)\overset{\lambda\to\infty}{\to}0
\end{aligned}
$$

Where,

$$
\begin{aligned}
&(n-s)\left(1-\left(\frac{R_N}{(n-s)}+\frac{R_C}{(n-s)}+\frac{R_D}{(n-s)}\right)+\ln\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)\right) \\
&\approx (n-s)\left(1-\left(\frac{R_N}{(n-s)}+\frac{R_C}{(n-s)}+\frac{R_D}{(n-s)}\right)-\left(1-\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)\right)\right. \\
&\left.-\frac{1}{2}\left(1-\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)\right)^2\right) \\
&=-R_N-\frac{(n-s)}{2}\left(1-\left(\frac{R_D}{(n-s)}+\frac{R_C}{(n-s)}\right)\right)^2 \\
&\approx -R_N-\frac{\eta^2}{2}
\end{aligned}
$$

(remark: for the last approximation see details in C.13)

Combining the expressions for C and D we get

$$E[W] = C + D = \frac{1}{\mu s}\frac{G}{A+B} + \frac{1}{\mu s}\frac{B}{A+B} = \frac{1}{\mu s}\frac{G+B}{A+B} = \frac{1}{\mu s}\frac{Ge^{-(R_N+R_D+R_C)}+\zeta}{\xi+\zeta}$$

where $\zeta = \zeta_1 - \zeta_2$. Thus, using the above approximation of $Ge^{-(R_N+R_D+R_C)}$ (C.16), we get

$$\begin{aligned}
\sqrt{s}E[W] &= \frac{1}{\mu\sqrt{s}}\frac{\sqrt{s}\left(\zeta_1\frac{1}{\beta} + \zeta_2\left(\frac{R_C+R_D}{R_N}\beta - \eta\sqrt{\frac{R_C+R_D}{R_N}} - \frac{1}{\beta}\right)\right) + \zeta_1 - \zeta_2}{\xi + \zeta_1 - \zeta_2} \\
&= \frac{\zeta_1\frac{1}{\beta} + \zeta_2\left(\frac{R_C+R_D}{R_N}\beta - \eta\sqrt{\frac{R_C+R_D}{R_N}} - \frac{1}{\beta}\right)}{\mu(\xi+\zeta_1-\zeta_2)} + \frac{(\zeta_1-\zeta_2)}{\sqrt{s}\mu(\xi+\zeta_1-\zeta_2)} \\
&\overset{s\to\infty}{\to} \frac{\frac{\phi(\beta)\Phi(\eta)}{\beta}\frac{1}{\beta} + \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\left(\frac{R_C+R_D}{R_N}\beta - \eta\sqrt{\frac{R_C+R_D}{R_N}} - \frac{1}{\beta}\right)}{\mu\left(\int_{-\infty}^{\beta}\Phi\left(\eta + (\beta-t)\sqrt{\frac{R_N}{R_C+R_D}}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)\right)}.
\end{aligned}$$

The approximations to ζ_1,ζ_2 and ξ are stated in (14.6), (14.7) and (14.8), respectively.

□

The next theorem gives the approximation for the case where $\beta = 0$.

Theorem 11. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Then*

$$\lim_{\lambda\to\infty}\sqrt{s}E[W] = \frac{1}{2\mu}\frac{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}\left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right)}{\sqrt{2\pi}\int_{-\infty}^{0}\Phi\left(\eta - t\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t) + \sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}\left(\eta\Phi(\eta)+\phi(\eta)\right)}$$

where $\eta_1 = \eta - \beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}$.

Proof. As before,

$$E[W] = \frac{1}{\mu s}\frac{G+B}{A+B}$$

We need to approximate G when $\beta = 0$.

$$G = \frac{(R_N)^s}{s!}\sum_{k=0}^{n-s-1}\sum_{j=0}^{n-s-k-1}\frac{1}{j!k!}(R_D)^j(R_C)^k\sum_{i=0}^{n-s-j-k-1} i\,(\rho)^i.$$

Using the formula

$$\sum_{l=0}^{M} l\rho^l = \frac{(M+1)(M)}{2} \tag{C.17}$$

one can show that when $\beta \to 0$, $\rho \to 1$ but still $\rho \neq 1$, the sum used in formula (C.15) is approximately equal to the one stated in (C.17).

Thus, using the fact that $s = R_N$, we get

$$G = \frac{(R_N)^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k \sum_{i=0}^{n-s-j-k-1} i (\rho)^i$$
$$= \frac{1}{s^{-s}s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k \frac{(n-s-j-k)(n-s-j-k-1)}{2}$$

using Stirling's formula, and Lemma 4 leading to

$$Ge^{-(R_N+R_C+R_D)} =$$
$$= e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k \frac{(n-s-j-k)(n-s-j-k-1)}{2}$$
$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$
$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{(j+k)}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$
$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$
$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{l=0}^{n-s-1} \sum_{j=0}^{l} \frac{l}{j!(l-j)!} (R_D)^j (R_C)^{l-j} (n-s-l)$$
$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$
$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{l=0}^{n-s-1} \frac{l}{l!} (n-s-l) \sum_{j=0}^{l} \frac{l!}{j!(l-j)!} (R_D)^j (R_C)^{l-j}$$
$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$
$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{l=0}^{n-s-1} \frac{l}{l!} (n-s-l) (R_D+R_C)^l$$
$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$
$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} (R_D+R_C) \sum_{l=0}^{n-s-2} \frac{1}{l!} (n-s-l-1) (R_D+R_C)^l$$

$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$

$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} (R_D+R_C) \sum_{l=0}^{n-s-1} \frac{1}{l!} (n-s-l-1) (R_D+R_C)^l$$

$$= \frac{(n-s-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$

$$- \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} (R_D+R_C) \sum_{l=0}^{n-s-1} \frac{1}{l!} (n-s-l) (R_D+R_C)^l$$

$$+ \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} (R_D+R_C) \sum_{l=0}^{n-s-1} \frac{1}{l!} (R_D+R_C)^l$$

$$= \frac{(n-s-R_D-R_C-1)}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} \sum_{k=0}^{n-s-1} \sum_{j=0}^{n-s-k-1} \frac{1}{j!k!} (R_D)^j (R_C)^k (n-s-j-k)$$

$$+ \frac{1}{2} e^{-(R_N+R_C+R_D)} \frac{R_N^s}{s!} (R_D+R_C) \sum_{l=0}^{n-s-1} \frac{1}{l!} (R_D+R_C)^l$$

$$= \frac{(n-s-R_D-R_C-1)}{2} \zeta + \frac{1}{2} e^{-R_N} \frac{R_N^s}{s!} (R_D+R_C) \sum_{l=0}^{n-s-1} \frac{1}{l!} (R_D+R_C)^l e^{-(R_C+R_D)}$$

$$\approx \frac{\eta\sqrt{R_D+R_C}}{2} \zeta + \frac{R_D+R_C}{2} e^{s-R_N} (\rho)^s \frac{1}{\sqrt{2\pi s}} \Phi(\eta)$$

$$\approx \frac{\eta\sqrt{R_D+R_C}}{2} \sqrt{\frac{R_C+R_D}{R_N}} \frac{1}{\sqrt{2\pi}} (\eta\Phi(\eta)+\phi(\eta)) + \frac{R_D+R_C}{2\sqrt{R_N}} (\rho)^s \frac{1}{\sqrt{2\pi}} \Phi(\eta)$$

$$\approx \frac{1}{\sqrt{2\pi}} \frac{R_D+R_C}{2\sqrt{R_N}} \left(\eta^2\Phi(\eta)+\eta\phi(\eta)\right) + \frac{1}{\sqrt{2\pi}} \frac{R_D+R_C}{2\sqrt{R_N}} (\rho)^s \Phi(\eta)$$

$$\approx \frac{1}{\sqrt{2\pi}} \frac{R_D+R_C}{2\sqrt{R_N}} \left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right) = \frac{\sqrt{s}}{2\sqrt{2\pi}} \frac{R_D+R_C}{R_N} \left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right).$$

Therefore, using Lemmas 3 and 4 we get

$$\sqrt{s}E[W] = \frac{1}{\mu\sqrt{s}} \frac{G+B}{A+B} = \frac{1}{\mu\sqrt{s}} \frac{\frac{\sqrt{s}}{2\sqrt{2\pi}} \frac{R_D+R_C}{R_N} \left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right) + \zeta}{\xi+\zeta}$$

$$\overset{s\to\infty}{\to} \frac{1}{2\sqrt{2\pi}} \frac{\frac{R_D+R_C}{R_N} \left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right)}{\mu\left(\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{R_N}{R_C+R_D}}\right) d\Phi(t) + \sqrt{\frac{R_C+R_D}{R_N}} \frac{1}{\sqrt{2\pi}} (\eta\Phi(\eta)+\phi(\eta))\right)}$$

$$= \frac{1}{2\mu} \frac{\frac{R_D+R_C}{R_N} \left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right)}{\sqrt{2\pi}\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{R_N}{R_C+R_D}}\right) d\Phi(t) + \sqrt{\frac{R_C+R_D}{R_N}} (\eta\Phi(\eta)+\phi(\eta))}$$

$$= \frac{1}{2\mu} \frac{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma} \left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right)}{\sqrt{2\pi}\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right) d\Phi(t) + \sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} (\eta\Phi(\eta)+\phi(\eta))}.$$

□

C.4 Proof of Approximation of the Probability of Blocking

In this appendix we will prove the approximation for the probability of blocking, stated in Section 14.3. The accurate measure was defined in Section 12.1, by Formula (12.1).

Theorem 12. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED conditions. Define $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, then*

$$\lim_{\lambda\to\infty}\sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2) + \phi(\sqrt{\eta^2+\beta^2})e^{\frac{\eta_1^2}{2}}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{B}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1)}$$

where $\eta_1 = \eta - \frac{\beta}{\sqrt{B}}$, $\nu = \frac{1}{\sqrt{1+B^{-1}}}$, $\nu_1 = \frac{\eta\sqrt{B^{-1}}+\beta}{\sqrt{1+B^{-1}}}$, $\nu_2 = \frac{\beta\sqrt{B^{-1}}-\eta}{\sqrt{1+B^{-1}}}$.

Proof. It follows from (12.1) that the probability of blocking is given by

$$\begin{aligned}
P_n &= \pi_0\left(\sum_{i=0}^{s}\sum_{j=0}^{n-i}\frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j\frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}\right.\\
&\quad\left.+\sum_{i=s+1}^{n}\sum_{j=0}^{n-i}\frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j\frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}\right)\\
&= \frac{\tilde{C}_1+\tilde{C}_2}{\tilde{A}+\tilde{B}_1-\tilde{B}_2}\cdot\frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}}{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}} = \frac{\delta_1+\delta_2}{\tilde{\xi}+\tilde{\zeta}_1-\tilde{\zeta}_2},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{C}_1 &= \sum_{i=0}^{s}\sum_{j=0}^{n-i}\frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j\frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}\\
\tilde{C}_2 &= \sum_{i=s+1}^{n}\sum_{j=0}^{n-i}\frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j\frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}\\
\tilde{A} &= \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n}}\frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\left(\frac{p\lambda}{(1-p)\delta}\right)^j\left(\frac{\lambda}{\gamma}\right)^k\\
\tilde{B}_1 &= \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s\frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^l\\
\tilde{B}_2 &= \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s\frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)^l
\end{aligned}$$

and

$$\tilde{\delta_1} = \sum_{i=0}^{s}\sum_{j=0}^{n-i} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j} e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$$

$$\tilde{\delta_2} = \sum_{i=s+1}^{n}\sum_{j=0}^{n-i} \frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j} e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$$

$$\tilde{\xi} = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n}} \frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\gamma}\right)^k e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$$

$$\tilde{\zeta_1} = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^l e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$$

$$\tilde{\zeta_2} = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho}+\frac{\lambda}{\gamma\rho}\right)^l e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}.$$

Note that by Lemmas 1,2 and 3

$$\lim_{\lambda\to\infty}\tilde{\zeta_1} = \lim_{\lambda\to\infty}\zeta_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta},$$

$$\lim_{\lambda\to\infty}\tilde{\zeta_2} = \lim_{\lambda\to\infty}\zeta_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1),$$

$$\lim_{\lambda\to\infty}\tilde{\xi} = \lim_{\lambda\to\infty}\xi = \int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t),$$

$$\begin{aligned}
\tilde{\delta_1} &= \sum_{i=0}^{s}\sum_{j=0}^{n-i} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j} e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)} \\
&= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}}{n!}\sum_{i=0}^{s}\sum_{j=0}^{n-i}\frac{n!}{i!j!(n-i-j)!}\left(\frac{\lambda}{(1-p)\mu}\right)^i\left(\frac{p\lambda}{(1-p)\delta}\right)^j\left(\frac{\lambda}{\gamma}\right)^{n-i-j} \\
&= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^n}{n!}\cdot \\
&\sum_{i=0}^{s}\sum_{j=0}^{n-i}\frac{n!}{i!j!(n-i-j)!}\left(\frac{\frac{\lambda}{(1-p)\mu}}{\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}\right)^i\left(\frac{\frac{p\lambda}{(1-p)\delta}}{\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}\right)^j\left(\frac{\frac{\lambda}{\gamma}}{\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}\right)^{n-i-j} \\
&= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^n}{n!}\sum_{i=0}^{s}\sum_{j=0}^{n-i}P(X_\lambda=(i,j,n-i-j)) \\
&= P(Y_\lambda=n)\sum_{i=0}^{s}\sum_{j=0}^{n-i}P(X_\lambda=(i,j,n-i-j)) \\
&= P(Y_\lambda=n)P(X_\lambda^1\le s)
\end{aligned}$$

where X_λ is a random variable with Multinomial distribution with parameters (n,p_i,p_j,p_k), $p_i = \frac{\frac{\lambda}{(1-p)\mu}}{\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}$, $p_j = \frac{\frac{p\lambda}{(1-p)\delta}}{\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}$, $p_k = \frac{\frac{\lambda}{\gamma}}{\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}$, Y_λ is a random variable with Pois-

son distribution with parameter $\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}$, and X_λ^1 is a random variable with Binomial distribution with parameters (n, p_i). By the CLT and the use of C.3

$$
\begin{aligned}
P(X_\lambda^1 \leq s) &= \Phi\left(\frac{s - np_i}{\sqrt{np_i(1-p_i)}}\right) = \Phi\left(\frac{s - n\frac{R_N}{R_N+R_C+R_D}}{\sqrt{n\frac{R_N}{R_N+R_C+R_D}(1-\frac{R_N}{R_N+R_C+R_D})}}\right) \\
&= \Phi\left(\frac{s - n\frac{R_N}{R_N+R_C+R_D}}{\sqrt{n\frac{R_N}{R_N+R_C+R_D}(\frac{R_C+R_D}{R_N+R_C+R_D})}}\right) = \Phi\left(\frac{s(R_N+R_C+R_D) - nR_N}{\sqrt{nR_N(R_C+R_D)}}\right) \\
&= \Phi\left(\sqrt{\frac{R_N}{R_C+R_D}}\frac{s\frac{R_N+R_C+R_D}{R_N} - n}{\sqrt{n}}\right) = \Phi\left(\sqrt{\frac{R_N}{R_C+R_D}}\frac{s(1+\frac{R_C+R_D}{R_N}) - n}{\sqrt{n}}\right) \\
&= \Phi\left(\sqrt{\frac{R_N}{R_C+R_D}}\frac{s+\frac{R_C+R_D}{\rho} - n}{\sqrt{n}}\right) = \Phi\left(-\sqrt{\frac{R_N}{R_C+R_D}}\sqrt{\frac{R_C+R_D}{n\rho}}\cdot\frac{n-s-\frac{R_C+R_D}{\rho}}{\sqrt{\frac{R_C+R_D}{\rho}}}\right) \\
&= \Phi\left(-\sqrt{\frac{s}{n}}\cdot\frac{n-s-\frac{R_C+R_D}{\rho}}{\sqrt{\frac{R_C+R_D}{\rho}}}\right) \approx \Phi\left(-\sqrt{\frac{R_N}{R_N+R_C+R_D}}\cdot\left(\eta - \beta\frac{R_C+R_D}{R_N}\right)\right) \\
&= \Phi\left(\frac{\beta\frac{R_C+R_D}{R_N} - \eta}{\sqrt{1+\frac{R_C+R_D}{R_N}}}\right) = \Phi\left(\frac{\beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} - \eta}{\sqrt{1+\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\right).
\end{aligned}
\tag{C.18}
$$

By the normal approximation of the Poisson distribution:

$$
\begin{aligned}
P(Y_\lambda = n) &\approx \frac{1}{\sqrt{\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}}\phi\left(\frac{n - \left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}}\right) \\
&\approx \frac{1}{\sqrt{s}\sqrt{1+\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\phi\left(\frac{\eta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} + \beta}{\sqrt{1+\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\right).
\end{aligned}
\tag{C.19}
$$

We based on the following equivalences (as λ tends to ∞) to develop Equations C.18 and C.19:

$$R_N + R_C + R_D = \frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} \approx s + (1-p)s\mu\left(\frac{p}{(1-p)\delta} + \frac{1}{\gamma}\right)$$
$$= s\left(1 + \frac{\mu\left(p\gamma + (1-p)\delta\right)}{\delta\gamma}\right) = s\left(1 + \frac{R_C + R_D}{R_N}\right);$$
$$n - \left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right) \approx s + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma} + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} - \left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)$$
$$= s + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} - \frac{\lambda}{(1-p)\mu} \approx s + \eta\sqrt{\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}} - s + \beta\sqrt{s}$$
$$\approx \eta\sqrt{\frac{s\mu\left(p\gamma + (1-p)\delta\right)}{\delta\gamma}} + \beta\sqrt{s};$$
$$\frac{n - \left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}{\sqrt{\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}}} \approx \frac{\eta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} + \beta}{\sqrt{1 + \frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}.$$

Following Equations C.18 and C.19 we get

$$\tilde{\delta_1} = P(Y_\lambda = n)P(X_\lambda^1 \leq s) \approx \frac{1}{\sqrt{s}\sqrt{1 + \frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\phi\left(\frac{\eta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} + \beta}{\sqrt{1 + \frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}} - \eta}{\sqrt{1 + \frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\right). \tag{C.20}$$

Now lets find an approximation for $\tilde{\delta_2}$.

$$\tilde{\delta_2} = \sum_{i=s+1}^{n}\sum_{j=0}^{n-i}\frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i\frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j\frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}$$
$$= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}}{s!s^{-s}}\sum_{i=s+1}^{n}\left(\frac{\lambda}{s(1-p)\mu}\right)^i\sum_{j=0}^{n-i}\frac{1}{j!(n-i-j)!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j\left(\frac{\lambda}{\gamma}\right)^{n-i-j}$$
$$= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}}{s!s^{-s}}\sum_{i=s+1}^{n}\left(\frac{\lambda}{s(1-p)\mu}\right)^i\frac{1}{(n-i)!}\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^{n-i}$$
$$= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}}{s!s^{-s}}\left(\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)^n\sum_{j=0}^{n-s-1}\frac{1}{j!}\left(\frac{\frac{\lambda}{s(1-p)\mu}}{\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}\right)^{n-j}$$
$$= \frac{e^{-\left(\frac{\lambda}{(1-p)\mu}+\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}\right)}}{s!s^{-s}}\left(\frac{\lambda}{s(1-p)\mu}\right)^n\sum_{j=0}^{n-s-1}\frac{1}{j!}\left(\frac{\frac{p\lambda}{(1-p)\delta}+\frac{\lambda}{\gamma}}{\frac{\lambda}{s(1-p)\mu}}\right)^{j}.$$

When comparing $\tilde{\delta_2}$ to ζ_2 form Equation C.2, we observe that

$$\tilde{\delta_2} = (1-\rho)\zeta_2 \approx \frac{\beta}{\sqrt{s}}\zeta_2.$$

Therefore, based on the approximation of ζ_2 from Lemma 2 we get

$$lim_{\lambda\to\infty}\tilde{\delta_2} = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\sqrt{s}}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1).$$

This proves Theorem 12. □

The next theorem gives the approximation for the case where $\beta = 0$.

Theorem 13. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED_0 conditions. Define and $B = \frac{R_N}{R_C+R_D} = \frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}$, then*

$$\lim_{\lambda\to\infty} \sqrt{s}P(block) = \frac{\nu\phi(\nu_1)\Phi(\nu_2) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{B}\right) d\Phi(t) + \frac{1}{\sqrt{B}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta) + \phi(\eta)\right)}$$

where $\nu = \frac{1}{\sqrt{1+B^{-1}}}$, $\nu_1 = \frac{\eta}{\sqrt{1+B}}$, $\nu_2 = \frac{-\eta}{\sqrt{1+B^{-1}}}$.

Proof. It follows from (12.1) that the probability of blocking is given by

$$P_n = \frac{\tilde{C_1} + \tilde{C_2}}{\tilde{A} + \tilde{B}} \cdot \frac{e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}}{e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}} = \frac{\tilde{\delta_1} + \tilde{\delta_2}}{\tilde{\xi} + \tilde{\zeta}},$$

where

$$\tilde{C_1} = \sum_{i=0}^{s}\sum_{j=0}^{n-i} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}$$

$$\tilde{C_2} = \sum_{i=s+1}^{n}\sum_{j=0}^{n-i} \frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j}$$

$$\tilde{A} = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n}} \frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\gamma}\right)^k$$

$$\tilde{B} = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l$$
$$- \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho} + \frac{\lambda}{\gamma\rho}\right)^l$$

and

$$\tilde{\delta_1} = \sum_{i=0}^{s}\sum_{j=0}^{n-i} \frac{1}{i!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j} e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}$$

$$\tilde{\delta_2} = \sum_{i=s+1}^{n}\sum_{j=0}^{n-i} \frac{1}{s!s^{i-s}}\left(\frac{\lambda}{(1-p)\mu}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{(1-p)\delta}\right)^j \frac{1}{(n-i-j)!}\left(\frac{\lambda}{\gamma}\right)^{n-i-j} e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}$$

$$\tilde{\xi} = \sum_{\substack{i,j,k|i\le s,\\ i+j+k\le n}} \frac{1}{i!j!k!}\left(\frac{\lambda}{(1-p)\mu}\right)^i \left(\frac{p\lambda}{(1-p)\delta}\right)^j \left(\frac{\lambda}{\gamma}\right)^k e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}$$

$$\tilde{\zeta} = \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)^l e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}$$
$$- \frac{1}{s!}\left(\frac{\lambda}{(1-p)\mu}\right)^s \frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{p\lambda}{(1-p)\delta\rho} + \frac{\lambda}{\gamma\rho}\right)^l e^{-\left(\frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} + \frac{\lambda}{\gamma}\right)}.$$

Note that by Lemmas 3 and 4

$$\lim_{\lambda\to\infty}\tilde{\xi}=\lim_{\lambda\to\infty}\xi=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}\right)d\Phi(t)$$

$$\lim_{\lambda\to\infty}\tilde{\zeta}=\lim_{\lambda\to\infty}\zeta=\sqrt{\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right),$$

In addition, the approximations for δ_1 and δ_2 are the same as the proof of Theorem 12.

$$\lim_{\beta\to0}\lim_{\lambda\to\infty}\sqrt{s}\delta_2=\frac{1}{\sqrt{2\pi}}\Phi(\eta)$$

$$\lim_{\beta\to0}\lim_{\lambda\to\infty}\sqrt{s}\delta_2=\frac{1}{\sqrt{1+\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\phi\left(\frac{\eta}{\sqrt{1+\frac{\delta\gamma}{\mu(p\gamma+(1-p)\delta)}}}\right)\Phi\left(-\frac{\eta}{\sqrt{1+\frac{\mu(p\gamma+(1-p)\delta)}{\delta\gamma}}}\right),$$

This proves Theorem 13.

□

References

[1] Trendwatch chartbook 2007 – trend affecting hospitals and health systems. American Hospital Association and Avalere, April 2007. 1.2

[2] The 2007 state of americas hospitals – taking the pulse. American Hospital Association, July 2007. 1.2

[3] Z. Aksin, M. Armony, and V. Mehrotra. The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665688, 2007. 2

[4] S. Anavi-Isakow and B. Golany. Managing multi-project environments through constant work-in-process. *International Journal of Project Management*, 21(1):9–18, 2003. 10.1

[5] M. Armony, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, and G. Yom-Tov. From emergency department to hospitalization: Using simulation, empirical and theoretical models for the operational analysis of ed, iw, and their interface. *Working Paper*, 2010. 7

[6] J.R. Artalejo and A. Gomez-Corral. *Retrial Queueing Systems (A Computational Approach)*. Springer, 2008. 2

[7] J. Atlason, M.A. Epelman, and S.G. Henderson. Optimizing call center staffing using simulation and analytic center cutting-plane methods. *MANAGEMENT SCIENCE*, 54(2):295–309, 2008. 9

[8] J.F. Bard and H.W. Purnomo. Hospital-wide reactive scheduling of nurses with preference considerations. *IIE Transactions*, 37(7):589–608, 2005. 10.3.3

[9] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, 1975. 31.1

[10] R. Bekker and A.M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, June 2009. DOI 10.1007/s10479-009-0570-z. 3.1, 27

[11] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004. 10.4

[12] J.W. Boudreau, W. Hopp, J.O. McClain, and L.J. Thomas. On the interface between operations and human resources management. *Manufacturing and Service Operations Management*, 5(3):179–202, 2003. 30

[13] E.K. Burke, P. de Causmaecker, G.V. Berghe, and H. Van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7(6):441–499, November 2004. 10.3.3

[14] Z. Carmon, J.G. Shanthikumar, and T.F. Carmon. A psychological perspective on service segmentation models: The significance of accounting for consumers' perceptions of waiting and service. *Management Science*, 41(11):1806–1815, 1995. 31.2

[15] E. Cedrá, L. de Pablos, and M.V. Rodríguezuría. Waiting list for surgery. In R.W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, chapter 6. Springer, 2006. 1.1

[16] B. Cheang, H. Li, A. Lim, and B. Rodrigues. Nurse rostering problems – a bibliographic survey. *European Journal of Operational Research*, 151(3):447–460, 2003. 10.3.3

[17] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, 2001. 12.2, 2, 19.1.1, A.1

[18] A.M. de Bruin, R. Bekker, L. van Zanten, and G.M. Koole. Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research*, 2009. 10.3.2, 28

[19] A.M. de Bruin, A.C. van Rossum, M.C. Visser, and G.M. Koole. Modeling the emergency cardiac inpatient flow. an application of queuing theory. *Health Care Management Science*, 10(2):125–137, June 2009. 1.1, 10.3.2

[20] K.C. Diwas and C. Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009. 7, 26.1

[21] S.G. Eick, W.A. Massey, and W. Whitt. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252, 1993. 3.3, 5.3.2

[22] S.G. Eick, W.A. Massey, and W. Whitt. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742, 1993. 3.3, 5.2

[23] A.K. Erlang. On the rational determination of the number of circuits. In E. Brockmeyer, H.L. Halstrom, and A. Jensen, editors, *The Life and Works of A.K. Erlang*. The Copenhagen Telephone Company, Copenhagen, Denmark, 1948. 3.2

[24] A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153(1):3–27, 2004. 10.3.3

[25] D.M. Fatovich and R.L. Hirsch. Entry overload, emergency department overcrowding, and ambulance bypass. *Emergency Medicine Journal*, 20:406–409, 2003. 1.1

[26] Z. Feldman, A. Mandelbaum, W.A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, February 2008. 2, 2.2, 3.1, 3.3, 5, 6.1, 6.1

[27] M.C. Fu. Optimization via simulation: A review. *Journal Annals of Operations Research*, 53(1):199–247, 1994. 9

[28] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: a tutorial and literature review. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003. Invited review paper. 1.1, 3.2

[29] P. Gonzales and C. Herrero. Optimal sharing of surgical costs in the presence of queues. *Mathematical Methods of Operations Research*, 59:435–446, 2004. 5

[30] W.J. Gordon and G.F. Newell. Cyclic queueing networks with exponential servers. *Operations Research*, 15(2):254–265, 1967. 11.2

[31] L. Green. How many hospital beds? *Inquiry - Blue Cross and Blue Shield Association*, 39(4):400–412, winter 2002/2003. 4, 10.3.2

[32] L. Green. Capacity planning and management in hospitals. In M.L. Brandeau, F. Sainfort, and W.P. Pierskalla, editors, *Operation Research and Health Care: A Handbook of Methods and Applications*, pages 14–41. Kluwer Academic Publishers, London, 2004. 4

[33] L. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operation Research*, 49(4):549–564, July-August 2001. 3.3, 6.1

[34] L. Green, P.J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service systems. *Production and Operations Management*, 16(1):13–39, January-February 2007. 3.1, 6.1

[35] L. Green, J. Soares, J.F. Giglio, and R.A. Green. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, January 2006. 3.1, 3.3

[36] L. Green and N. Yankovic. A queueing model for nurse staffing. Working paper, 2007. 20.6

[37] Linda V. Green, Sergei Savin, and Mark Murray. Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety*, 33(4):211–218, April 2007. 1.1, 27

[38] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, 3 edition, 1998. 12.2

[39] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–587, 1981. 3.2, 4, 6, 6.1, 8

[40] A. Hall and G. Walton. Information overload within the health care system: a literature review. *Health Information and Libraries Journal*, 21:102–108, June 2004. 3

[41] R.W. Hall, editor. *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer, 2006. 1.1, 1, 1.1

[42] P.R. Harper and A.K. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of Operations Research Society*, 53:11–18, 2002. 10.3.2

[43] A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and B. Zwart. Refining square root safety staffing by expanding erlang-c. *to appear in Operations Research*, 2008. 3.2

[44] O. Jennings, A. Mandelbaum, W. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, October 1996. 2.2, 3.3

[45] O.B. Jennings and F. de Véricourt. Nurse-to-patient ratios in hospital staffing: a queueing perspective. Working Paper, Duke University, 2007. (document), 3, 1.1, 3.1, 3.2, 10.1, 10.3.3, 10.4, 29, 10.4

[46] O.B. Jennings and F. de Véricourt. Dimensioning large-scale membership services. *Operations Research*, 56(1):173–187, January 2008. (document), 2, 10.4, 29, 10.4, 16.1

[47] S.A. Jones, M.P. Joy, and J. Pearson. Forecasting demand of emergency care. *Health Care Management Science*, 5(4):297–305, November 2002. 10.3.2

[48] E.P.C. Kao and G.G. Tung. Forecasting demands for inpatient services in a large public health care delivery system. *Socio-Economic Planning Sciences*, 14(2):97–106, 1980. 10.3.2

[49] P. Khudyakov. Designing a call center with an IVR. Master's thesis, Technion - Israel Institute of Technology, 2006. 10.1, 16.2, 17, 19.1.1

[50] P. Khudyakov, P. Feigin, and A. Mandelbaum. Designing a call center with an IVR (interactive voice response). Under revision to QUESTA, 2009. 1

[51] A. Lahiri and A. Seidmann. Analyzing the differential impact of radiology information systems across radiology modalities. *Journal of the American College of Radiology*, 6(10):522–526, October 2009. 2.1

[52] S. Lundgren and K. Segesten. Nurses use of time in a medical-surgical ward with all-rn staffing. *Journal of Nursing management*, 9:13–20, 2001. 20.6

[53] A. Mandelbaum, W. Massey, and M. Reiman. Strong approximations for markovian service networks. *Queueing Systems*, 30(1-2):149–201, November 1998. 2, 5.1, 8, A.2, B, B, B, B.3

[54] A. Mandelbaum, W.A. Massey, M. Reiman, and B. Rider. Time varying multiserver queues with abandonment and retrials. In P. Key and D. Smith, editors, *ITC-16, Teletraffic Engineering in a Competitive World*, pages 355–364. Elsevier, 1999. 8

[55] A. Mandelbaum, W.A. Massey, M. Reiman, A. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2-4):149–171, December 2002. 8, B.2.1

[56] Y.N. Marmor. *Emergency-Departments Simulation in Support of Service-Engineering: Staffing, Design, and Real-Time Tracking*. PhD thesis, Technion - Israeli Institute of Technology, February 2010. 22, 23, 26.1

[57] Y.N. Marmor and A. Mandelbaum. Emergency department (ed) data at 5 anonymous hospitals. Technical report, Technion - Israeli Institute of Technology, 2007. 15

[58] Y.N. Marmor and D.A. Sinreich. Emergency departments operations: the basis for developing a simulation tool. *IIE Transactions*, 37(3):233–245, 2005. 1.1, 7

[59] W. Massey and W. Whitt. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems*, 13:183–250, 1993. 5

[60] W. Massey and W. Whitt. An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *Annals of applied Probability*, 4(4):1145–1160, 1994. 6

[61] H.E. Miller, W.P. Pierskalla, and G.J. Rath. Nurse scheduling using mathematical programming. *Operations Research*, 24(5):857–870, Sep. - Oct. 1976. Special Issue on Health Care. 7

[62] Israel Ministry of Health. Financial report for years 2000-2005. http://www.health.gov.il/pages/default.asp?maincat=1&catid=98&pageid=4078, November 2006. 1.2, 3.1

[63] California Department of Health Services (CDHS). Nurse-to-patient staffing ratio regulations. http://www.dhs.ca.gov/Inc/NTP/defealt.htm, January 2004. 1.1, 3.1, 10.3.3

[64] A.D. Polyanin and A.V. Manzhirov. *Handbook of Integral Equations*. CRC Press, Boca Raton, 2nd edition, 2008. 5

[65] A. Puhalskii. On the invariance principle for the first passage time. *Mathmatics of Opererations Research*, 19(4):946–954, November 1994. B.2, B.4

[66] R.S. Randhawa and S. Kumar. Multi-server loss systems with subscribers. *Mathematics of Operations Research*, 34(1):142–179, February 2006. 10.4

[67] D. Sitompul and SU. Randhawa. Nurse scheduling models: a state-of-the-art review. *J Soc Health Syst*, 2(1):62–72, Spring 1990. 3.1, 10.3.3

[68] V.L. Smith-Daniels, S.B. Schweikhart, and D.E. Smith-Daniels. Capacity management in health care services: Review and future research directions. *Decision Sciences*, 19(4):889–919, December 1988. 10.3.1

[69] B. Sobolov, A. Levy, and L. Kuramoto. Access to surgery and medical consequences of delays. In R.W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, chapter 3. Springer, 2006. 1.1

[70] M.L. Spearmana, D.L. Woodruffa, and W.J. Hoppa. CONWIP: a pull alternative to kanban. *International Journal of Production Research*, 28(5):879–894, 1990. 10.1

[71] Y. Tseytlin. Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments. Master's thesis, Technion - Israel Institute of Technology, April 2009. 2, 22, 28

[72] W. Whitt. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics*, 55(5):476–484, 2007. 5.2.2

[73] G. Yom-Tov and A. Mandelbaum. Queues in hospitals: Semi-open queueing networks in the QED regime. Technical report, Technion - Israeli Institute of Technology, 2008. 3.1

[74] S. Zeltyn, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, Y.N. Marmor, A. Mandelbaum, A. Shtub, T. Lauterman D. Schwartz, K. Moskovitch, S. Tzafrir, and F. Basis. Simulation-based models of emergency departments: Operational, tactical and strategic staffing. *Submitted to ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 2009. 3.1, 3.3