A QUEUEING MODEL OF HOSPITAL CONGESTION

by

Pouya Bastani

B.Sc., Simon Fraser University, 2007

a Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the Department
of
Mathematics

SIMON FRASER UNIVERSITY
Summer 2009

APPROVAL

Name:	Pouya Bastani
Degree:	Master of Science
Title of Thesis:	A queueing model of hospital congestion

Examining Committee: Dr. Paul Tupper
Chair

Dr. Ralf Wittenberg
Senior Supervisor

Dr. Sandy Rutherford
Co-supervisor

Dr. Rachel Altman
SFU Examiner

Date Approved: June 25, 2009

Declaration of Partial Copyright Licence

Abstract

Over the years, the growing population in British Columbia has led to escalating wait times and overcrowding in hospital Emergency Departments (EDs), due to insufficient number of beds in specific units of the hospitals, such as the Intensive Care Unit (ICU) or the Medical Unit (MU). To enhance the level of access to care, a successful prediction of bed requirements is needed. This is achieved by having an adequate model of the patient flows to and between the different compartments of the hospital. Focusing only on the stream of emergency patients, we developed a queueing network to model the interaction between the ICU and the MU, which is believed to be causing a major proportion of the congestion in the ED. Through approximate analytical methods and simulation, we determined sufficient bed counts in each of these two units so as to guarantee certain access standards.

Acknowledgments

I would like to thank my supervisors Dr. Sandy Rutherford and Dr. Ralf Wittenberg for their teachings and guidance in this thesis. I would also like to thank the IRMACS Centre for the administrative and technical support that greatly facilitated this research. Finally, I am grateful to my parents, Bijan and Mahnaz, and my girlfriend, Helen, for their care and continual support throughout my graduate studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Acute care is the treatment of a severe medical condition for only a short period of time and at a crisis level. Many hospitals are acute care facilities with the goal of discharging the patient as soon as the patient is deemed healthy and stable. The rising population in BC has created a need to understand better how hospital resources relate to the quality of service in acute care facilities in British Columbia. In this thesis, hospital resources are measured in beds, a term we use to refer to the physical count of locations able to provide in-patient services, accompanied by the necessary equipment and staff. To ensure an adequate level of access to care, it is important to examine future bed requirements. The accurate prediction of this count requires both the knowledge of future population demographics, which affects the demand for acute care services, and also an understanding of how the number of available beds affects access to care.

This work is dedicated to the latter issue. More specifically, our goal is to understand how the flow of patients to, within the different compartments of, and out of the hospital affects access to care. This would enable us to to estimate the required number of beds that would guarantee a certain access level. This is important, for a low hospital capacity leads to patients in need of care being turned away, and growing waiting lists cause stress on other hospital units. For example, when insufficient medical beds are available to meet demand, emergency medical patients spill over into surgical beds; consequently, surgical waiting lists increase as planned admissions are postponed. Determining bed requirements is also important from the hospital management perspective, for it has direct implications on staff allocations and operation costs. For the purposes of this project, we consider only bed requirements for patients arriving through the emergency department; inclusion of the elective stream of patients into the model is left as a future task. Hereafter, we will refer to inpatients simply as patients; outpatients are not considered here, for they are not hospitalized overnight, and thus do not affect bed requirements.

The focus of this project is on understanding and quantifying the blocking phenomenon that occurs in the Intensive Care Unit when patients who need to be transferred out of this unit to another cannot find a free bed. This congestion, in turn, causes delay in providing beds to newly arrived critically ill patients. In the next section, we describe the hospital units considered in this project, followed by the definition of the access measure that we use here. Finally, we state the goal of this project and describe its relation to the past work done at the Complex Systems Modelling Group at IRMACS.

1.1 Hospital Units

Acute care hospitals are divided into multiple compartments that may differ from one facility to another, depending on the size and location of the hospital. To keep the model general, we consider the following three units, which we believe can be applied to most acute care hospitals:

- **Emergency Department (ED)**: sometimes termed Emergency Room, this unit provides initial treatment to patients with a broad spectrum of illnesses and injuries, some of which may be life-threatening and requiring immediate attention.

  The process from patient arrival in the ED to placement in a bed can be summarized as follows: It begins with the triage nurse, who determines the urgency of the patient's condition. Next, the patient is seen by an ED physician, who, after possible diagnostic testing, determines whether or not the patient requires admission to the hospital. In some cases, after the initial assessment and treatment, patients are discharged or transferred to another hospital for various reasons. Otherwise, a bed is requested in the appropriate nursing unit (e.g., medical, surgical, or intensive care). The availability of a bed is affected not only by the capacity of the relevant unit, but also by the admission and scheduling policies of elective patients, particularly surgical patients who compete for the same beds. If a bed is not available readily, the patient is kept in the ED and is treated by a nurse. This may result in the discharge of the patient directly from the ED if the treatment is completed before a bed becomes available. Otherwise, when a bed is reported to be free in the required unit, the patient is transferred to the bed.

- **Intensive Care Unit (ICU):** is a specialized department used for intensive care medicine. Most patients arriving to the ICU are admitted from the ED. After their treatment, ICU patients are usually transfered to the medical unit for further care before discharge. This transfer, however, is possible only if a bed is available in the medical unit.

- **Medical Unit (MU):** is a term used in this project to refer to the rest of the hospital. This unit is designated for patients from the ED whose severity of illness is not sufficiently high to be considered for the ICU. In addition, patients from the ICU spend some time in the MU for full recovery before their complete discharge from the hospital. Finally, scheduled patients (not considered here) are taken to this unit. When patients' treatment is completed in the MU, they leave the hospital or are assigned a bed in the Alternative Level of Care (ALC) unit of the MU, depending on their health condition. The ALC, which shares beds with the MU, is essentially a waiting unit for patients to be transferred to a residential care unit when these institutions are fully occupied.

1.2 Access to Care

Access to care can be measured in several different manners. Traditionally, hospital bed capacity decisions have been made based on Target Occupancy Rate (TOR) – the average percentage of occupied beds – and the most commonly used occupancy target has been 85%. Another metric often cited in the literature is the Target Access Rate (TAR), which measures the percentage of the time that a census count will show that the hospital contains at least one empty bed.

The measure that we use here for emergency patients, developed in collaboration with the B.C. Ministry of Health Services, is the Target Time to Access (TTA), which is the targeted percentage of times that patients receive beds within a given maximal time delay. For example, a TTA of 80% within 6 hours implies that there are enough beds to keep the percentage of the patients that have to wait longer than 6 hours below 20%. We will use TTA in this project for the purpose of determining bed requirements for emergency patients.

1.3 Project Goal

This project emerged as part of my work with Complex Systems Modelling Group (CSMG) at IRMACS, and its goal has been to study more fundamentally some of the aspects of the queueing network that has been developed to model patient flows in B.C. acute care hospitals.

In the first phase of the acute care modelling project at CSMG [31], most hospitals were divided into 5 subunits: ICU, Pediatrics, Psychiatry, Surgical, and Medical. These units were assumed to be operating independently of one another. The project addressed the issue of access to care for emergency patients with the assumption that elective admissions have higher priority. Although it may seem counterintuitive, it is reasonable to assume that patients admitted through the ED have lower priority than elective admissions, since the former group, upon finding the hospital full, are placed in the ED where they receive care, while the latter group would probably face a cancellation if a bed is not available as scheduled. Given the historical admission rates through the ED, the model related TTA to the number of beds in the hospital. For instance, for any one hospital, the model determined the required number of beds in the pediatrics unit so that a TTA of 85% within 6 hours was achieved. In that project, however, the process by which elective admissions could get cancelled due to hospital overcrowding was not considered; this task was undertaken in the second phase of the project [5]. To model the cancellation mechanism, it was assumed that even though scheduled patients had priority over emergency patients, if they waited longer than a certain amount, their appointments would be cancelled.

In this project, I have attempted to study a more realistic model of the hospital, by developing a queueing network that takes into account the transfer of patients from the ICU to other hospital units, which we collectively refer to as the Medical Unit (MU). The main issue that arises in this setting is the possibility of the MU being full, which causes patients whose treatment is completed in the ICU to be blocked from being transferred, and hence to have to stay in the ICU until a bed becomes available in the MU. This effectively lengthens their duration of stay in the ICU, and reduces the hospital efficiency, as the impact of blocking in the ICU propagates to the ED, where some patients may be waiting to get into the ICU. The blocking mechanism is studied in this project both numerically and also via approximate methods. However, as both methods fail to completely encompass the overall model due to its complexity, discrete-event simulation is used to understand how the number of beds affect TTA for emergency patients.

Besides blocking, another queueing principle that is investigated in this project is reneging, or the departure of units[1] from the system before having completed service. In most of the literature, reneging is due to the impatience of customers who are in the queue, and so it is often viewed as customers leaving the queue. In this project, however, reneging is proposed as a model both for deaths that occur in the hospital (mainly the ED and the ICU), and also for the treatment completions that take place in the ED, which result in patients leaving the hospital before receiving a bed. Although the underlying reason for these departures is not impatience, they can be treated similarly. However, the point that must be emphasized is that deaths that occur in the ICU are analogous to customers leaving the service station while they are being served. This is not what is commonly referred to as reneging. On the other hand, deaths in the ED or direct discharges from the ED, which is viewed as the queue to the hospital, conform with the widely known notion of reneging. In both cases, a large number of reneged patients indicates low quality of care, since high death rates and direct departures from the ED often occur when the hospital is operating at or near full capacity. The study of reneging in this project is hoped to give us an understanding of how such quantities as the percentage of reneged patients are affected by the number of beds in the hospital.

[1] In queueing theory, the term "units" refers to customers arriving at a service station. In the hospital model developed here, the term is used to refer to patients arriving at the ED.

Since the ideas we use in our modelling approach lie in the domain of queueing theory, the next chapter is devoted to the fundamentals of this subject, including notation, terminologies, and important results used in this thesis. In the following chapter, we review some of the past work done in applying queueing models in a health care setting. We will then give a concrete description of the queueing model that we have developed to describe the interaction between the ED, ICU, and MU. After studying reneging in chapter 5 and tandem queues with blocking in chapter 6, we will apply the results in chapter 7 to find an estimate for the required number of beds in the ICU and the MU, for the purpose of achieving a TTA of 90% within 1 hour for the ICU and 80% within 6 hours for the MU. This estimate is then improved upon using the simulation software SimEvents in MATLAB.

Chapter 2

Fundamentals and New Results

Queueing theory is a mathematical approach in Operations Research applied to the analysis of waiting lines. A.K. Erlang first analyzed queues in 1913 in the context of telephone facilities. The body of knowledge that developed thereafter via further research and analysis came to be known as Queueing Theory, and is extensively applied in industrial settings and retail sectors. The use of queueing theory and other principles of operations management in health care is fairly recent, with applications which can often be thought of as a balance between

(i) minimizing costs due to occupation of resources such as beds; and

(ii) minimizing wait times of patients.

In effect, Target Time to Access is a measure that achieves this balance by incorporating what health care officials believe is a compromise between cost and wait time minimization. In general, the analysis of queueing systems consists of evaluating a set of performance measures, such as mean customer wait time or mean server idle time (customers and servers, respectively, represent patients and beds in the queueing model of hospitals developed later).

Queueing systems are often analyzed by analytical methods or simulation. The latter is a general technique of wide application able to incorporate many complexities of a model, but its main drawback is the potentially high development and computational cost to obtain accurate results. Analytical methods, on the other hand, can often produce results in relatively short time, but often require that the model satisfy a more restrictive set of assumptions and constraints in order to make the derivation possible. When deriving analytical solutions becomes intractable, numerical solutions of the underlying equations (see the next section) may also be considered, but even this approach is limited in its application, because the memory required to store the state space of queueing networks grows exponentially with the number of service stations.

Although the results in this project are extensively based on simulation and numerical solutions, better understanding of the model developed here can be obtained through analytical investigations. However, a full study of the whole model consisting of the interactions between the ED, the ICU, and the MU is essentially impossible using analytical techniques. Nevertheless, imposing simplifying assumptions and using approximate techniques, useful insights can be gained. In this section, we now give a short introduction to some basics of queueing theory, which will be used in subsequent chapters.

2.1 Rate Matrix

Let $\{X_n, n \geq 0\}$ be a Markov chain over a finite state space $\mathcal{S} = \{0, 1, 2, \ldots, N\}$. Define p_{ij} as the probability of transition from state i to j in one step, i.e.

$$p_{ij} = \Pr\{X_n = j \big| X_{n-1} = i\},$$

where we assume that the chain is homogeneous so that p_{ij} is independent of n (usually n denotes discrete points in time, in which case this is equivalent to requiring that $\mathcal{S}$ be the set of states at equilibrium, assuming the equilibrium exists). The matrix $P = (p_{ij})$ for $i, j \in \mathcal{S}$ is called the *transition probability matrix* of the Markov chain. This matrix has the property that $\sum_{j \in \mathcal{S}} p_{ij} = 1$, since the probability of transitioning from state i to some state in $\mathcal{S}$ must be 1.

Now, let the row vector π be the equilibrium probability distribution with elements $\pi_i = \Pr\{X_n = i\}$. This vector must remain unchanged under the application of the transition matrix; in other words, it is defined as the eigenvector of the probability matrix associated with the eigenvalue 1 so that

$$\pi P = \pi. \tag{2.1}$$

Assuming the equilibrium exists, the *rate matrix*[1] can be defined as

$$Q = P - I. \tag{2.2}$$

Then, equation (2.1) can be written as

$$\pi Q = 0 \tag{2.3}$$

where 0 is the zero vector. The normalization condition $\sum_{i \in \mathcal{S}} \pi_i = 1$ can be incorporated in equation (2.3) by replacing the elements in the last column of Q with 1's (call this new matrix $\tilde{Q}$) and replacing the last element of the zero vector on the right hand side with 1 (call this new vector b). With these definitions, the equilibrium distribution π satisfies the following equation:

$$\pi \tilde{Q} = b. \tag{2.4}$$

The rate matrix $\tilde{Q}$ is often quite sparse, because each state can only transition to a small number of neighboring states. Various efficient numerical methods are available for solving such equations. MATLAB is especially efficient at recognizing the sparse structure of the rate matrix and hence applying specialized techniques to dramatically reduce the cost of solving equation (2.4) when the state space is large, as is the case in our queueing network model of the hospital. In section 6.1 we demonstrate the use of the rate matrix in obtaining the queue length distribution of the tandem queueing system discussed with blocking.

2.2 Characteristics of Queueing Systems

Conceptually, the simplest queueing model is the single server queue illustrated in figure 2.1 (often the waiting space itself is not illustrated in the diagram, and its presence is implied, unless otherwise stated). The system models the flow of customers as they arrive, wait in the queue if the server is busy, receive service, and eventually leave.

[1] In analyzing the transient period, when states and transition probabilities are time dependent, the rate matrix gives the rate at which the state vector $\pi(t)$ changes with time via the following equation

$$\pi(t)' = \pi(t)Q(t).$$

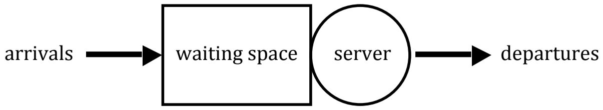

Figure 2.1: The single server queue

A queueing system consists of customers who have a certain arrival pattern, and are served at a station consisting of a number of servers with a specific service pattern. In this respect, we can see that a basic queueing system, one that consists of a single service station, can be described by the following characteristics:

(i) **arrival pattern:** This is specified by the distribution of interarrival time of customers. An important related quantity is the mean interarrival time. The reciprocal of the mean interarrival time is referred to as the arrival rate. A commonly used distribution for the interarrival time of customers is the exponential distribution (see section 2.4), which is determined by the mean alone. Though it can be otherwise, we consider only arrivals that occur singly (not in batches).

(ii) **service pattern:** It is specified by the distribution of the time taken to complete service. The reciprocal of the mean service time is referred to as the service rate. As with the arrival pattern, the service pattern is commonly described by the exponential distribution. We assume departures occur one by one, and not in batches.

(iii) **number of servers:** A number of servers may work in parallel, and an arriving unit can choose randomly between any of the free servers. If all servers are busy, the unit joins a queue common to all the servers.

(iv) **system capacity:** There might be situations in which a queueing system can only accommodate a limited number of waiting units. In this case, if the number of waiting customers plus those in service exceeds the system capacity, any further arrival does not join the system and is lost.

(v) **queue discipline:** If a customer arrives at the system at a time when the server(s) is (are) unavailable to provide service, he/she is forced to wait in the queue temporarily. If there is more than one customer waiting in the queue at a time the server becomes available, one of the customers in the queue is selected to start receiving service. The manner in which waiting customers are taken in for service when a new server becomes available is referred to as the service discipline. Throughout this thesis First-Come, First-Served (FCFS) is assumed as the service discipline.

We next introduce a notation in queueing theory that is used to describe single-station queues in a short format.

2.3 Kendall's Notation

A basic queueing system can often be described by a notation introduced by Kendall. Referring to the numbering used above in section 2.2, this notation takes the form (i)/(ii)/(iii)/(iv) so that, for example, a queueing system with exponential interarrival and service time distribution, c servers, and system capacity k, is represented by $M/M/c/k$, where M stands for Markovian. Unless otherwise mentioned, the service discipline is assumed to be FCFS. Moreover, if the waiting capacity is infinite, i.e. $k = \infty$, the last symbol may be omitted, so that the notation for the above example would simply become $M/M/c$.

2.4 The Exponential Distribution

The simplest queueing models assume that the interarrival and service times are exponentially distributed, so that, for instance, if λ is the mean arrival rate, then the probability density function (pdf) for the time between successive arrivals would be

$$f(t) = \lambda e^{-\lambda t}. \tag{2.5}$$

Equivalently, the arrivals can be said to follow the Poisson process, a collection $\{N(t), t \geq 0\}$ of random variables, where $N(t)$ is the number of events that have occurred up to time t, starting from time 0. The Poisson distribution is given by

$$Pr\{N(t) = n, t \geq 0\} = \frac{(\lambda t)^n e^{-\lambda t}}{n!}. \tag{2.6}$$

We now state three important properties of the exponential and Poisson distributions that we use in this thesis:

(P1) *Memoryless Property*: If X is an exponentially distributed random variable, then

$$\Pr\{X \geq x + y | X \geq x\} = \Pr\{X \geq y\}. \tag{2.7}$$

This property has the following implication: if service times are exponentially distributed, then the probability that a customer's service is completed at some future time is independent of how long the customer has already been in service. It is mainly because of this special property that the exponential distribution has been the most widely used distribution in the analysis of queueing systems.

(P2) *Additive Property*: The sum of n independent Poisson processes with parameter λ_i, for $i = 1, 2, \ldots, n$, is a Poisson process with parameter $\lambda_1 + \lambda_2 + \cdots + \lambda_n$.

(P3) *Decomposition Property*: Suppose that $N(t)$ is a Poisson process with rate λ and that each arrival is marked with probability p independent of all other arrivals. Let $N_1(t)$ and $N_2(t)$ respectively denote the number of marked and unmarked arrivals in $[0, t]$. Then $N_1(t)$ and $N_2(t)$ are two independent Poisson processes with respective rates λp and $\lambda(1 - p)$.

2.5 Little's Theorem

For a queueing system at equilibrium with arrival rate λ, mean queue length L, and mean wait time W, Little's Theorem states

$$L = \lambda W. \tag{2.8}$$

The profoundness of this formula is due to the fact that it holds for virtually all queueing systems under very general conditions. Furthermore, the same relation holds if L and W represent the mean number of units and mean wait time in the system[2] at any time point, respectively. Although the initial insight into the truth of this relation is due to Morse [25], it was his student Little [23] who gave the rigorous proof of the formula.

[2] The term "system" is used to refer to both service station and queue in combination.

2.6 Relationship between Wait Time and Queue Length

We now derive an equation that relates the equilibrium queue length distribution of an $M/G/c$ queue to the equilibrium distribution of the wait time in the queue (G stands for the general distribution); as far as we are aware, this result has not been published elsewhere. To do so, we first state Burke's Theorem and the PASTA property.

Let us first define the following quantities:

$$a_n = \Pr\{\text{an arrival finds } n \text{ units in the queue}\}$$

$$d_n = \Pr\{\text{a departure leaves } n \text{ units in the queue}\}$$

$$q_n = \Pr\{\text{a random observer finds } n \text{ units in the queue}\}$$

Then *Burke's Theorem* states that for any queueing system at equilibrium in which arrivals and departures occur one by one (no batch arrivals or departures) it must be true that

$$a_n = d_n. \tag{2.9}$$

On the other hand, the PASTA (Poisson Arrivals See Time Averages) property states that in any queueing system in which the arrivals follow a Poisson process,

$$a_n = q_n. \tag{2.10}$$

Thus, for a queueing system at equilibrium with Poisson arrivals in which both arrivals and departures occur individually, we must have that

$$d_n = q_n\,. \tag{2.11}$$

Let $w(t)$ be the probability density function (pdf) of the wait time in the queue[3]. Equation (2.11) can then be used to find a formula relating the probability generating function (pgf) of the queue distribution

$$P(z) = \sum_n q_n z^n \qquad |z| < 1 \tag{2.12}$$

[3] Throughout this thesis, we use wait time to refer to the time spent in the queue, not in the system. To refer to the latter, we will specifically state wait time in the system.

to the Laplace Transform (LT) of the wait time pdf

$$w^*(s) = \int_0^\infty e^{-st}\, w(t)\, dt \tag{2.13}$$

for an $M/G/c$ queue with mean arrival rate λ. To do so, first note that in a queue with FCFS service discipline, the probability that there are n units in the queue when a unit leaves the queue (and enters the service station) is equal to the probability that n units arrive during its wait time. This leads to the following expression:

$$d_n = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, w(t)\, dt. \tag{2.14}$$

Using (2.11) the pgf of the queue length distribution can be written as:

$$\begin{aligned} P(z) &= \sum_{n=0}^{\infty} q_n z^n \qquad\qquad |z| < 1 \\ &= \sum_{n=0}^{\infty} d_n z^n \\ &= \sum_{n=0}^{\infty} \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} z^n\, w(t)\, dt. \end{aligned} \tag{2.15}$$

Now, as stated by Widder [36, p. 446], if the Laplace transform

$$\int_0^\infty e^{-st}\phi(t)\, dt$$

converges absolutely at a point $s = s_0$, then for $\mathrm{Re}\{s_0\} < \mathrm{Re}\{s\} \le \mathrm{Re}\{R\}$ it must converge uniformly, where R is an arbitrary complex number. Since $P(z)$ is finite for all $0 < \mathrm{Re}\{z\} < 1$, the integral in (2.15) must be convergent. Moreover, because the integrand is non-negative, the integral is absolutely convergent, and by the above statement, it must be uniformly convergent. This allows us to interchange the order of integration and summation, or more precisely, the order of the limits in

$$\sum_{n=0}^{\infty} \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} z^n\, w(t)\, dt = \lim_{N\to\infty}\ \lim_{y\to\infty} \sum_{n=0}^{N} \int_0^y \frac{e^{-\lambda t}(\lambda t)^n}{n!} z^n\, w(t)\, dt,$$

so that expression (2.15) can further be simplified:

$$\begin{aligned} P(z) &= \int_0^\infty w(t)\, e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(z\lambda t)^n}{n!}\, dt \\ &= \int_0^\infty w(t) e^{-(1-z)\lambda t}\, dt \\ &= w^*[(1-z)\lambda]\,. \end{aligned} \tag{2.16}$$

Alternatively, defining $s = (1-z)\lambda$, we can write equation (2.16) as

$$w^*(s) = P(1 - s/\lambda). \tag{2.17}$$

Therefore, knowing the pgf of the queue length distribution, we can obtain the wait time distribution by inverting the LT. Little's Theorem can also be obtained from this expression:

$$W = -\frac{d}{ds} w^*(s)\Big|_{s=0} = \frac{1}{\lambda} P'(1) = \frac{L}{\lambda}. \tag{2.18}$$

It should be mentioned that a result similar to equation (2.17) was derived previously for the $M/G/1$ queue: Let $F(z)$ be the pgf of the number of units in the *system*, and let $v^*(s)$ be the LT of the distribution of wait times in the *system*. Then, according to Gross and Harris [16],

$$v^*(s) = F(1 - s/\lambda). \tag{2.19}$$

By comparison, our result (2.17) holds for the more general case of the $M/G/c$ queue, but only when considering the pgf of the queue length and the LT of the queue wait time. To see why equation (2.19) cannot be applied to an $M/G/c$ queue, note that equation (2.14) holds only for that part of the system in which no unit that has arrived later than another can leave earlier. For an $M/G/c$ queueing system, this can only be guaranteed for the queue portion of the system. Moreover, formula (2.17) cannot be applied to queueing systems with reneging, which refers to situations in which units can leave the queue without receiving service; for this reason, in section 5 we shall explicitly need to derive the wait time distribution for the $M/M/c$ queue with reneging, and cannot rely on the knowledge of the queue distribution alone.

Note also that equations (2.17) and (2.19) assume the analytical continuation of $P(z)$ for values of $\mathrm{Re}\{z\} < -1$. To see this, consider the formula for inversion of LT using the Fourier series method of evaluating the Bromwich contour integral [1]

$$w(t) = \frac{2e^{at}}{\pi} \int_0^\infty \mathrm{Re}\{w^*(a+iu)\} \cos(ut)\, du, \tag{2.20}$$

where a is chosen such that $w^*(s)$ has no singularities on or to the right of $s = a$. It is clear that $w^*(s)$ needs to be calculated for very large values of $|s|^2 = a^2 + u^2$. In view of equation (2.17) this implies that $P(z)$ must be evaluated for very large negative values of $\mathrm{Re}\{z\}$, for which the sum $\sum_{n=0}^{\infty} q_n z^n$ used in defining $P(z)$ may not be convergent. Thus we use analytical continuation to define $P(z)$ for $\mathrm{Re}\{z\} < -1$. As a result, equation (2.17) is well-defined and can be inverted.

In the absence of an analytical formula for the pdf of the queue length, such as when we compute the probabilities numerically using the rate matrix, we can only compute $P(z)$ numerically from the series

$$P(z_i) = \sum_{m=0}^{M} q_m z_i^m, \tag{2.21}$$

where $|z_i| < 1$ and M is the maximum value of m for which q_m is computed. Without a functional expression for $P(z)$, inverting the LT of $w^*(s)$ becomes more challenging. To overcome this difficulty, we first approximate $P(z)$ by a rational polynomial, say $\tilde{P}(z)$. In our work we used the function `ratpolyfit(z,P,kn,kd)` implemented by Godfrey [13] in MATLAB, which, given values of a function $P(z)$ at points z_i, finds two polynomials $N(z)$ and $D(z)$ of orders k_n and k_d, respectively, such that the rational polynomial $\tilde{P}(z) = N(z)/D(z)$ best approximates $P(z)$ in the least squares sense, so that the error

$$e_0 = \sum_i \left|\tilde{P}(z_i) - P(z_i)\right|^2 \tag{2.22}$$

is minimized. The motivation behind using a rational approximation to the queue length pgf is that the $M/M/c$ queue has an exact rational queue length pgf (see equation (2.35)).

Let us now write the rational approximation $\tilde{w}^*(s) = \tilde{P}(1 - s/\lambda)$ to $w^*(s)$ as follows:

$$\tilde{w}^*(s) = c_0 + \frac{R(s)}{D(s)}, \tag{2.23}$$

where $R(s)$ is another polynomial with degree smaller than k_d. We can easily find the inverse LT of this expression to obtain the approximate wait time distribution $\tilde{w}(t)$ as

$$\tilde{w}(t) = c_0\delta(t) + \sum_{k=1}^{k_d} c_k e^{-r_k t}\,, \tag{2.24}$$

where r_k is the kth pole among the k_d poles of $\tilde{w}^*(s)$, which has a residue of c_k at $s = r_k$. Here, c_0 is an approximation to the probability of not having to wait in the queue at all, which, in the case of the $M/G/c$ queue, is equal to the probability that not all servers are occupied, i.e. $c_0 \approx p_0 + p_1 + \cdots + p_{c-1}$. Note that if an equilibrium is to exist, this probability is always non-zero, i.e. $c_0 > 0$. Hence, expression (2.23) implies that the degree of the numerator must be equal to that of the denominator in the rational approximation $\tilde{P}(z)$; in other words, we must choose $k_n = k_d \equiv k$.

For every particular problem, we choose k experimentally by starting from $k = 1$, and increasing it incrementally, comparing error estimates for different values of k. It must be mentioned that higher values of k do not necessarily yield smaller error as measured by (2.22), for as Godfrey [13] comments in the introduction to his code:

If you overfit the data, then you will usually have pole-zero cancellations and/or poles and zeros with a very large magnitude. If that happens, then reduce the values of k_n and/or k_d.

He further adds that the approximation becomes ill-conditioned with higher values of k_n and/or k_d. Thus, it is preferable to confine our search to small values of k. Clearly, measuring the error in the s-space ($z = 1 - s/\lambda$) as measured by (2.22) also gives a good indication of the accuracy of the approximation in the t-space. This is because the inversion of the transform $\tilde{w}^*(s)$ is highly sensitive to the location of its poles, and as Godfrey [13] mentions,

often, if you have a good fit, you will find that your polynomials have roots where the real function has zeros and poles.

Besides measuring the error in the s-space, we can look at the following two quantities as means of measuring the error in the t-space:

1) free server probability, and

2) integral of wait time distribution.

From what we already mentioned, the better the approximation, the smaller the difference

$$e_1 = \left| \sum_{n=0}^{c-1} q_n - c_0 \right|, \tag{2.25}$$

since c_0 estimates the free server probability. Furthermore, since every probability distribution must be normalized, the error

$$e_2 = \left| 1 - \int_0^\infty \tilde{w}(t)\, dt \right| \tag{2.26}$$

needs to be small. Therefore, together with e_0, these three error measures allow us to find a desirable value of k. In section 6.1 we demonstrate the method outlined here to compute wait time distribution in a tandem queueing system of finite intermediate waiting capacity.

2.7 The $M/M/c$ Queue

We now illustrate the ideas introduced in this chapter with the use of an example. Consider the $M/M/c$ queue where the arrival and service rates are λ and μ, respectively. Assuming that steady state exists, let p_n be the steady state distribution of the number of units in the system. We proceed to derive the equations involving p_n by using the *rate-equality principle*, which states that the rate at which a process enters a state is equal to the rate at which it leaves that state.

Consider state 0, when there are no units in the system. The process can leave this state only when there is an arrival, which causes the system to transition to state 1. The long-run proportion of time the process is in state 0 is p_0, and since λ is the rate of arrival, the rate at which the process leaves state 0 to go to state 1 is λp_0. Moreover, the process can enter state 0 only from state 1 through a departure or service completion. Since the proportion of time the process is in state 1 is p_1 and the rate of leaving state 1 through service completion is μ, the rate at which the process transitions from state 1 to 0 is μp_1. Using the rate-equality principle, we get

$$\lambda p_0 = \mu p_1. \tag{2.27}$$

Now consider state $0 < n < c$. The process can leave state n in two ways, either through an arrival or through a departure. The proportion of time the process is in state n is p_n and the total rate at which the process leaves state n through arrivals or departures is $\lambda p_n + n\mu p_n$, since there are n servers busy (additive property of the Poisson process). The process can enter state n in two ways, either through arrival from state $n-1$ or through a departure from state $n+1$. Thus, the rate at which the process enters state n is $\lambda p_{n-1} + \mu p_{n+1}$. By the rate-equality principle

$$\lambda p_n + n\mu p_n = \lambda p_{n-1} + (n+1)\mu p_{n+1}\,. \tag{2.28}$$

Similarly, for the case of $n \geq c$, we get

$$\lambda p_n + c\mu p_n = \lambda p_{n-1} + c\mu p_{n+1}\,. \tag{2.29}$$

Repeated application of (2.28) along with (2.27) at the last step yields

$$\begin{aligned} \lambda p_n - (n+1)\mu p_{n+1} &= \lambda p_{n-1} - n\mu p_n \\ &= \lambda p_{n-2} - (n-1)\mu p_{n-1} \\ &\vdots \\ &= \lambda p_0 - \mu p_1 \\ &= 0. \end{aligned}$$

By rearranging terms and iterating we obtain that for $0 < n \leq c$

$$p_n = \frac{\lambda/\mu}{n}\,p_{n-1} = \frac{(\lambda/\mu)^2}{n(n-1)}\,p_{n-2} = \cdots = \frac{(\lambda/\mu)^n}{n!}\,p_0. \tag{2.30}$$

In a similar fashion, we get that for $n > c$

$$p_n = \frac{(\lambda/\mu)^n}{c!c^{n-c}}\,p_0. \tag{2.31}$$

Now for $\lambda/(c\mu) < 1$, the normalization condition $\sum_{n=0}^{\infty} p_n = 1$ gives

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\lambda/c\mu)}\right]^{-1}. \tag{2.32}$$

We now proceed to compute some performance measures. The expected queue length L can be computed as

$$L = \sum_{n=c}^{\infty}(n-c)p_n = \frac{\lambda p_c}{\mu(1-\rho)^2}, \tag{2.33}$$

where $\rho = \lambda/c\mu$ is referred to as the server utilization. Applying Little's formula, we also obtain the expected waiting time in the queue

$$W = \frac{L}{\lambda} = \frac{p_c}{\mu(1-\rho)^2}. \tag{2.34}$$

Knowing the probability distribution, we can now directly compute the pgf of the number in the queue

$$P(z) = \sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} p_n z^{n-c} = 1 - \frac{p_c}{1-\rho} + \frac{p_c}{1-\rho z}, \tag{2.35}$$

which allows us to find the LT of the wait time distribution as

$$w^*(s) = P(1 - s/\lambda) = 1 - \frac{p_c}{1-\rho} + \frac{p_c}{1-\rho+s/c\mu}. \tag{2.36}$$

Inverting the transform gives

$$w(t) = \left(1 - \frac{p_c}{1-\rho}\right)\delta(t) + c\mu p_c\, e^{-c\mu(1-\rho)t} \qquad t > 0. \tag{2.37}$$

Note that the coefficient of $\delta(t)$ is the probability of zero wait, or the probability that there is a free server upon arrival. It is important to realize that computing p_c in this coefficient becomes numerically difficult when the number of servers c and the server load λ/μ are very large, as from (2.31) it can be seen that both $(\lambda/\mu)^c$ and $c!$ then become extremely large, introducing numerical errors. To address this issue Adan and Resing [2] first note that

$$\begin{aligned}\frac{p_c}{1-\rho} &= \frac{(c\rho)^c/c!}{(1-\rho)\sum_{n=0}^{c-1}(c\rho)^n/n! + (c\rho)^c/c!} \\ &= \frac{\rho B(c-1, c\rho)}{1-\rho+\rho B(c-1,c\rho)},\end{aligned}$$

where $B(c, \rho)$ is Erlang's B-formula given by

$$B(c,\rho) = \frac{\rho^c/c!}{\sum_{n=0}^{c-1}\rho^n/n! + \rho^c/c!}. \tag{2.38}$$

Then, by dividing the numerator and denominator by $\sum_{n=0}^{c-1}\rho_n/n!$, the authors obtain the following recursive relation for Erlang's B-formula:

$$B(c,\rho) = \frac{\rho B(c-1,\rho)}{c+\rho B(c-1,\rho)}, \tag{2.39}$$

which can be used to avoid division of large numbers in equation (2.37).

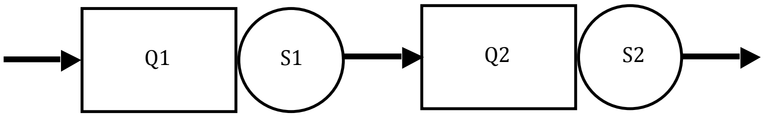

Figure 2.2: Two infinite-capacity queues in tandem

2.8 Tandem $M/M/c$ Queues

Now consider an $M/M/c_1$ queue placed in tandem with an $M/M/c_2$ queue, with an infinite queue allowed in between (Fig. 2.2). Customers discharged from the first station must proceed to the next to have their second phase of service completed. It can easily be shown that the equilibrium distribution

$$p_{mn} = \Pr \{ \ m \text{ units in } S_1 \text{ and } n \text{ units in } S_2 \ \}$$

has the product form

$$p_{mn} = p_1(m) \cdot p_2(n) \tag{2.40}$$

where $p_i(\cdot)$ is the distribution of the number of units in an $M/M/c_i$ queue. This shows that the two stations operate independently of one another; this is the typical behaviour of queueing networks with unlimited queueing space in between. An important theorem regarding the *output process* of the $M/M/c$ queue at equilibrium states that the inter-departure times are independently and identically distributed as an exponential random variable with mean $1/\lambda$, where λ is the arrival rate. It follows that if the arrival rate into S_1 is λ, then the arrival rate into S_2 is also λ. The general result applicable to networks of queues with infinite waiting capacity between each is known as the Jackson Theorem [18] .

Chapter 3

Literature Review

In the past, queueing theory has been effectively used in such areas of health care modelling as staff scheduling, policy making (for example, determining how prioritizing certain groups of patients affects wait times), and bed requirement analysis, which is the focus of this thesis.

It is common practice in health services to estimate the required number of beds as the average number of daily admissions times average length of stay in days and divided by average bed occupancy rate (average number of occupied beds during a day) [17]:

$$\text{bed requirement} = \frac{\text{average no. of daily admissions}}{\text{average bed occupancy rate}} \times \text{average length of stay.} \tag{3.1}$$

However, as de Bruin *et al.* mention in [8], "a model, only based on average numbers, is not capable of describing the complexity and dynamics of the in-patient flow." Moreover, reported occupancy levels are generally based on the average midnight census (for billing purposes), which results in underestimation of the bed requirements.

More recently, queueing models have provided better means of estimating the necessary number of beds based on sound performance measures. In [28], Pike *et al.* use the $M/G/\infty$ queue as a model for the casualty ward of a hospital. They show that in steady state, the bed occupancy rate follows a Poisson distribution with mean λW, where λ denotes the daily admission rate and W denotes the average duration of stay. Using this model, the authors determine the required number of beds in order to guarantee that a given target percentage of arrivals receive a bed immediately.

Weiss and McClain [35] also use the $M/G/\infty$ system to model the queue of patients needing alternative levels of care in acute care facilities whose treatment is completed and who are waiting to be transferred to an extended care facility (ECF). These patients are kept in the hospital due to unavailability of beds in the ECF and reduce the hospital utilization. The authors' model allows managers to predict the effect of certain policy changes on appropriate access measures. For instance, the cost-benefit trade-off of opening an additional extended care facility within a region is compared to that of assigning a higher priority to patients going to ECF from acute care facilities than to those coming from other sources.

Instead of using an infinite capacity queue, Worthington [37] uses an $M/G/c$ queue with a state-dependent arrival rate to address the long hospital-wait list problem. He experiments with various management actions such as increasing the number of beds or decreasing mean service times through appropriate means.

Gorunescu *et al.* [14] develop a queueing model for the movement of patients through a hospital department. Performance measures, such as mean bed occupancy and the probability of rejecting an arriving patient due to hospital overcrowding, are computed. These quantities enable hospital managers to determine the number of beds needed in order to keep the fraction of delays under a threshold, and also to optimize the average cost per day by balancing the costs of empty beds against those of delayed patients.

Although service times, unlike inter-arrival times, do not usually have an exponential distribution, such an assumption is often made in order to simplify the analysis greatly. For instance, de Bruin *et al.* [8] use the $M/M/c/c$ queue, referred to as the Erlang Loss model, to investigate the emergency in-patient flow of cardiac patients in a university medical centre in order to determine the optimal bed allocation so as to keep the fraction of refused admissions under a target limit. The authors find the relation between the size of a hospital unit, occupancy rate, and target admission rates. A cancellation rate of 5% is often considered acceptable. However, while the target occupancy rate of 85% has become a golden standard in health care [15], the authors note that using one target occupancy rate for

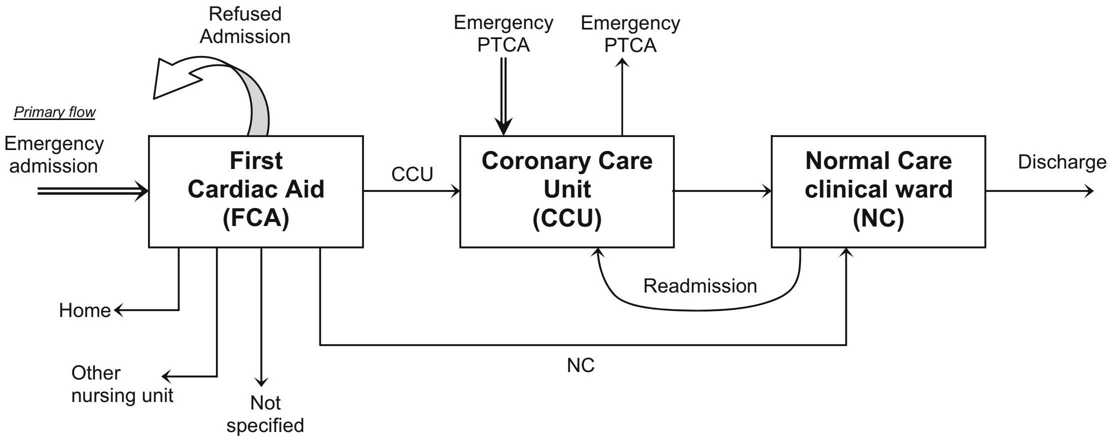

Figure 3.1: Flow of the emergency cardiac patients (from [8])

hospital units of different size is not reasonable, for larger hospitals can usually operate at a higher occupancy rate than smaller ones.

After analytically estimating the required number of beds in the First Cardiac Aid (FCA) unit of the medical centre, de Bruin *et al.* [8] also use numerical methods to determine the number of beds in the Coronary Care Unit (CCU) and the Normal Care clinical ward (NC), which are situated downstream from the FCA (Fig. 3.1). The authors had to rely on numerical techniques at this stage, because the finite capacity of the CCU and the NC leads to blocking in the FCA, making analytical calculations extremely difficult to carry out.

In fact, due to the complexities that arise in analyzing queueing systems with multiple interacting service stations, the study of health care facilities has mainly been done using simulation, with analytical methods applied to the study of one hospital as a whole (represented by a single service station) or of single hospital units, assumed to operate independently of the others. In recent years, however, approximate analytical methods have been developed and used in studying multi-facility interactions.

For instance, Koizumi *et al.* [20] use a queueing network model with blocking to model the congestion in mental health facilities in Philadelphia (Fig. 3.2). Their results point out that a shortage of a particular type of facilities could be the main cause of the blocking, which

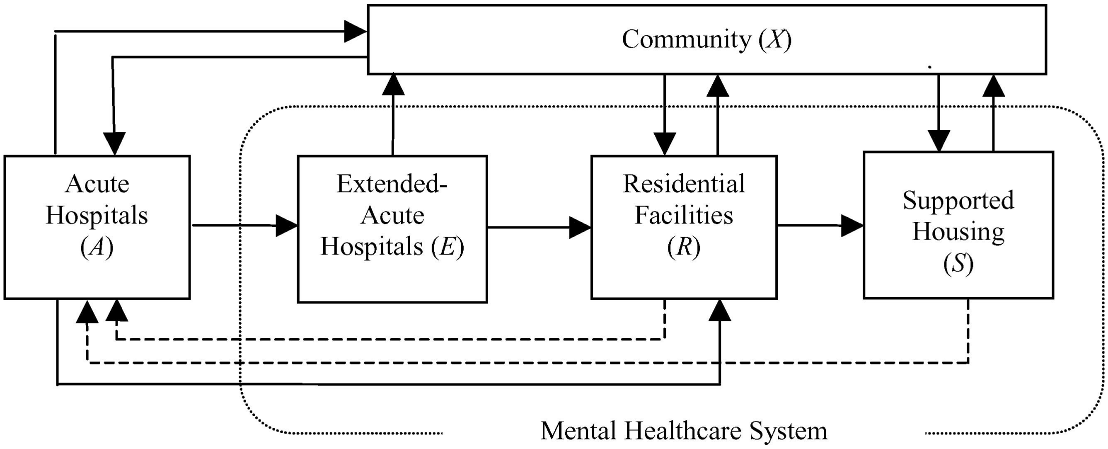

Figure 3.2: Patient flow between different facilities (from [20])

results in many patients spending unnecessary extra days in intensive care facilities. Their system consists of three types of psychiatric institutions:

(E) *extended acute hospitals*: designated to patients who require follow-up care after being discharged from an acute hospital

(R) *residential facilities*: accommodate those clients who require basic daily living support

(S) *supported housing*: provides clients with a minimum daily living support

(A) *acute hospitals*

(X) *all other*: composed of various accommodations for psychiatric patients

Due to the fact that stations E, R, and S have only a finite number of beds and no other waiting space, blocking may occur at the flows $E \rightarrow R$ and $R \rightarrow S$. In the absence of blocking, this queueing network could be decomposed into individual independent institutions, resulting in a product-form solution for the equilibrium distribution of the number of units in each unit. To incorporate blocking into this product-form solution as an approximation, Koizumi *et al.* use an effective service rate that is modified by the expected waiting time at the upstream stations. The authors note that their approximation only holds when the total number of beds at adjacent upstream stations is large enough to accommodate the steady-state number of waiting patients, a condition which is only known a posteriori.

Using their approximate analytical solution, they found that the system-wide congestion is

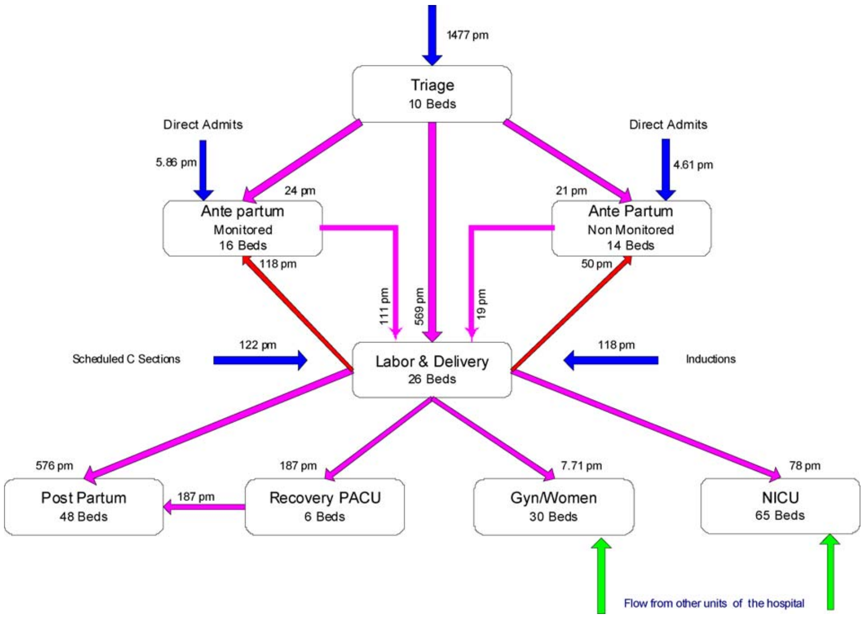

Figure 3.3: Queueing network model of an obstetrics hospital (from [9])

due primarily to shortages only in the supported housing facility, which causes blocking in the other facilities. Thus, a possible solution from the policy viewpoint is to consider an increase in the number of beds in this specific unit.

Another queueing network model applied to a hospital setting is that of Cochran and Bharti [9], who study a specific obstetrics hospital consisting of 8 subunits with 4 different patient arrival streams (Fig. 3.3). The transfer of patients between the different compartments creates blocking in some of the units.

As a precursor to building their simulation model, the authors first use an approximate analysis of the network by ignoring blocking and time dependence of the parameters. This helps to provide quick answers to many of the management questions, in addition to guiding them in validation of their simulation in special circumstances. By using discrete-event simulation to model the full interaction of the different subunits and patient groups, the authors then compare alternative methods of reducing blocking times and increasing the hospital throughput. For example, after identifying the Post Partum unit as the bottleneck of the system, they show that by adding new beds to this unit or reallocating beds from underutilized units, such as Medicine/Surgery, the maximum number of deliveries per month can be increased; however, in the latter case, reallocating beds beyond a certain threshold causes the bottleneck to shift to a different unit. An interesting result is that increasing the number of beds in the bottleneck unit by 15% yields a 38% improvement in the overall hospital throughput.

As can be seen, the application of queueing models to healthcare is growing more popular as hospital management teams are gaining awareness of the advantages of these operational research techniques in addressing such issues as determining optimal bed counts and making policy decisions with regards to resource allocation. Research in applying queueing networks with blocking is rarer in the literature due to the mathematical complexities involved in computing performance measures associated with such systems. As a result, hospitals with interacting subunits are often studied through simulations, for they are able to incorporate much more detail than is affordable by analytical methods. In this thesis, we use both approximate analytical techniques and simulation to study a simple queueing network composed of only two service stations placed in tandem. In the next section, we discuss the details of our model.

Chapter 4

Model Overview

Before delving into the details of the model developed in this project, it is worth reviewing the previous model [31] that was developed by the CSMG group at IRMACS to predict the hospital bed counts in B.C. Our current model is a step towards further understanding how the interaction of the Intensive Care Unit with other compartments of the hospitals affects the flow of patients. This interdependence was not considered in the previous projects of the Acute Care group at the CSMG, as it was assumed that all the different hospital units operate independently of one another.

4.1 Segmented Multi-Stream Model

In summary, in the first stage of the modelling of B.C. acute care hospitals by the CSMG, three streams of patients were considered:

- emergency
- elective direct
- direct transfers from other facilities

The primary focus of the project was to find a relationship between access to care and hospital bed counts. The access measure used in the project was a TTA of 80% within 6 hours; in other words, the goal was to find the least number of beds in each of the hospital compartments that guaranteed that 80% of the patients arriving in the ED receive a bed in their designated unit within 6 hours.

The model separated the hospitals into multiple units, which were assumed to operate independently of each other. Each compartment was given its own queue consisting of the three streams of patients. The emergency patients were given lower priority than the other two patient streams. In other words, if a bed became available while patients from all three units were waiting for that bed, then a patient from the ED was transferred to the bed only if there were no patients in the other two streams waiting for the bed. Amongst the elective and transferred patients, the decision was on a first-come first-served basis. While waiting, the emergency patients received treatment in the ED queue and could possibly leave the hospital directly; direct discharge from the ED is explained in more detail later.

In the next section, we will focus only on the stream of emergency patients, but will consider their transfer from the ICU to one of surgical and medical units, which we have combined into one entity, called the Medical Unit (MU).

4.2 The ED-ICU-MU Model

In this project, we look only at the stream of Emergency patients into the hospital. A more complete model would include the elective admissions and transfers from other facilities. However, our current model by itself involves the interaction of three hospital units (Fig. 4.1), and we felt it necessary to gain complete understanding of this simpler model before adding further complexity. The three hospital units considered in this project are the Emergency Department, the Intensive Care Unit, and the Medical Unit. Our aim is to understand how the flow of patients among these three units causes delays in obtaining beds.

Upon arrival in the ED, patients require a bed either in the ICU or the MU. If there is bed available in the required unit, the arriving patient goes there without waiting in the ED. Otherwise, the patient is kept in the ED and waits for his/her required unit. When a bed becomes available, the earliest of such patients is transferred to the specific unit. However, while the patient is kept in the ED, he/she is treated by nurses as necessary.

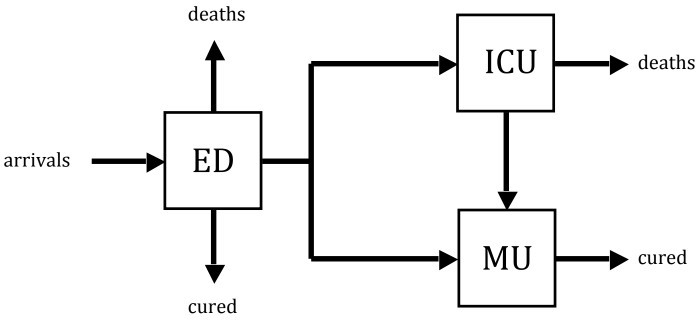

Figure 4.1: The interaction of ED, ICU, and MU

As a result of treatment in the ED, it is in fact possible in certain circumstances that a patient is deemed healthy enough to leave the hospital directly from the ED without requiring further stay. This usually applies to patients going to the MU. There are also patients who unfortunately die in the ED due to their long wait; those in this group most often are waiting for an ICU bed and have a critical health condition.

In queueing theory terminology, both groups of patients can be seen as reneging units[1], such that when their reneging time[2] is exceeded, they leave the queue (the ED). The event where a patient leaves the hospital directly from the ED, whether due to treatment completion or death, is referred to as Direct Discharge From Emergency (DDFE). A relatively large number of DDFEs is indicative of the fact that the hospital does not have enough beds to provide adequate access to care. Hence, it is desirable to keep the fraction of DDFEs very low. Note that in our model the ED is viewed as the queue to either the ICU or the MU, but one from which units (patients) may renege due to “impatience” (treatment completion or death). It must also be mentioned that the ED is assumed to be able to provide as many beds as necessary to treat the waiting patients. Thus, the ED can be viewed as an an infinite-capacity waiting space in which reneging is possible.

[1] Reneging refers to customers becoming impatient and leaving the queue without receiving service. A detailed treatment of an $M/M/c$ queue with reneging is given in chapter 5.

[2] Reneging time is the amount of time a unit is willing to wait before leaving the system.

While in the ICU, it is also possible that a patient may undergo a medical fatality. In such a case, the deceased individual is transferred out of the ICU bed (and hospital), thus allowing a patient from the ED to be transferred to the bed. We will refer to deaths in the ICU as reneging in service. In most cases of successful treatments in the ICU, however, the patient needs to be transferred to the MU.

This transfer to the MU is only possible if a bed is free there. In the event that a patient cannot be transferred, she/he is kept in the ICU. In other words, such patients are keeping the bed occupied, even though their required intensive care is completed. Note that patients waiting in the ICU cause the hospital efficiency to lower, as they block patients in the ED from receiving those beds. This phenomenon not only delays access to care, but it also generates some financial loss because the blocking patients in the ICU are ready to move to a less intensive and hence less expensive unit. Hence, controlling the congestion in this unit is important not only from a clinical perspective, but also from a budgetary perspective for health care policy makers.

When a bed becomes free in the MU, the doctor decides on whether to admit a patient from the ED or the ICU into the MU, if there are patients in both units waiting for a bed. In this project we assume that blocked ICU patients are given a higher priority for two reasons:

1. Keeping a patient in an ICU bed is more costly than placing her/him in an MU bed, especially since this patient no longer needs intensive care.
2. By freeing ICU beds quickly, the target access of 90% within 1 hour can more easily be achieved.

4.3 Modelling Assumptions

In almost every model there are certain assumptions that are used to simplify the modelling process, since otherwise, in an attempt to model every detail of the real world scenario, the model would become too complex to understand and useless for any practical purpose. Here we list the assumptions incorporated in our model:

(1) Inter-arrival times, service times, and reneging times are all exponentially distributed.

Moreover, reneging times are assumed to be independent of the time spent in queue.

(2) Rates of service, arrivals, and reneging, are all assumed to be constant in time. A more realistic model would include hourly variations in the arrival rate accompanied by overall seasonal or even day-of-week changes. On the other hand, because the lengths of stays are of the order of days, and average treatment times usually do not vary throughout the year, it is reasonable to assume that the service rate is constant.

(3) There is no delay in between patient transfers: if a patient leaves a bed, another waiting patient immediately replaces him/her, with no intermediate delay. In real hospital settings, some time is taken in cleaning and preparation for the next patient.

(4) On average, patients being transferred out of the ICU to the MU and those arriving directly from the ED require the same amount of treatment time in the MU.

(5) Blocked ICU patients have priority over those waiting in the ED to enter the MU.

(6) The amount of time spent in any one unit of the hospital (ED, ICU, or MU) is independent of the time spent in any other unit.

(7) The ED always has enough beds.

(8) Patients in the ED are served on the FCFS service discipline.

Furthermore, as mentioned before, we have ignored the elective direct and transferred patients in our model in this preliminary stage. A future model will incorporate all three streams of patients (see section 7.5).

Chapter 5

$M/M/c$ Queue with Reneging

Reneging refers to the situation in which customers waiting in line to be served become impatient and leave the queue. In our model, the ED is viewed as the queue, and there are two reasons for which emergency patients may leave the ED without getting a bed:

1. medical fatality
2. treatment completion

As mentioned in the previous section, only ICU patients are assumed to pass away in the ED due to their severe condition, while treatment completions can occur for those patients waiting for an MU bed. Also, recall that patients who are already in the ICU may die before their treatment is completed. This latter situation is referred to as in-service reneging.

The most common reneging mechanism, which we use here, is the one in which the units' maximum waiting times are exponentially distributed and independent of time in queue. We also assume here that the time spent in service is independent of the time spent in queue. In most systems, waiting in queue involves no service; however, in the hospital model considered here, patients receive treatment in the ED, and so a more realistic assumption would be to incorporate the queueing time in the overall service time.

Here we analyze the $M/M/c$ queue with reneging. As we shall see in chapter 6, the ICU and the MU may be approximated by two such queues, operating independently of each

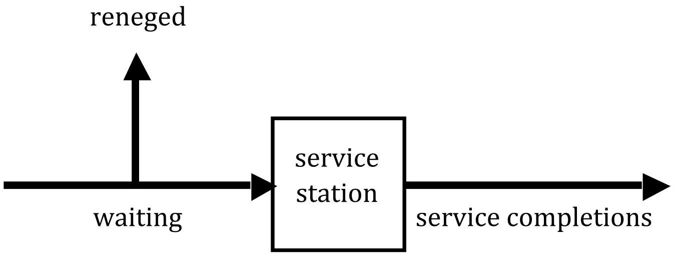

Figure 5.1: Reneging from queue

other. In addition to computing the wait time distribution, we derive a formula for the percentage of patients who renege, which can be used to obtain the mean reneging time. This is important because from the available data, we cannot directly obtain the reneging parameter, while the percentage of reneged units can easily be calculated.

5.1 Queue Distribution

Consider the $M/M/c$ queue, where units in the queue may leave when their reneging time, assumed to be exponentially distributed with parameter α, has expired. As usual, we denote the arrival rate by λ and service rate of each of the c servers by μ.

Ancker and Gafarian [3] studied the $M/M/1/N$ queue with exponential reneging time and balking. They considered a balking mechanism in which an arrival finding n units in the system leaves without joining the queue with probability n/N. They obtained an expression for the wait time distribution of the units that join the queue and successfully receive service without reneging. Thus, by taking the limit of the distribution as N approaches infinity, one can obtain the result for the $M/M/1$ queue with reneging only. This approach is presented in the Appendix, and the result is used as a check of the correctness of the formula obtained in this chapter for $c = 1$.

Due to reneging, this system always reaches equilibrium. To see this, consider a busier queue of type $M/M/\infty$ with arrival rate λ and service rate $\tilde{\mu} = \min(\alpha, \mu)$. This queue has an equilibrium for all values of $\lambda/\tilde{\mu} > 0$. It follows that the $M/M/c$ queue with reneging parameter α and service rate μ also has an equilibrium. Now, let p_n be the equilibrium probability distribution of the number of units in the system. As in section 2.7, by applying the rate-equality principle we arrive at the following set of equations:

$$n\mu p_n = \lambda p_{n-1} \qquad n \leq c, \tag{5.1}$$

$$[c\mu + \alpha(n-c)]p_n = \lambda p_{n-1} \qquad n > c. \tag{5.2}$$

Note that when $n > c$, the number of units in the system is $n - c$ so that the time until the next departure due to reneging is exponential with rate $(n-c)\alpha$. Combined with the service rate of the c servers, the departures from the system when all servers are busy is exponential with parameter $c\mu + \alpha(n-c)$ leading to the second equation.

By iterating equation (5.1) we obtain

$$\begin{aligned} p_n &= \frac{\lambda}{n\mu} p_{n-1} \qquad\qquad n \leq c \\ &= \frac{\lambda^2}{n(n-1)\mu} p_{n-2} \\ &\vdots \\ &= \frac{\lambda^n}{n!\mu^n} p_0. \end{aligned}$$

Similarly, for $n > c$ we can solve (5.2) in an iterative manner to get

$$\begin{aligned} p_n &= \frac{\lambda}{c\mu + (n-c)\alpha}\, p_{n-1} \qquad\qquad n > c \\ &\vdots \\ &= \frac{\lambda^{n-c}}{\prod_{j=1}^{n-c}(c\mu + j\alpha)}\, p_c \\ &= \frac{\lambda^n}{c!\mu^c \prod_{j=1}^{n-c}(c\mu + j\alpha)}\, p_0. \end{aligned}$$

Now, rescaling the arrival and service rate by the reneging parameter via

$$\delta = \lambda/\alpha \tag{5.3}$$

$$\gamma = \mu/\alpha \tag{5.4}$$

we can write the distribution of the number of units in the system as

$$p_n = \frac{\delta^n}{n!\gamma^n} p_0 \qquad n \leq c, \tag{5.5}$$

$$p_n = \frac{\delta^n}{c!\gamma^c \prod_{j=1}^{n-c}(c\gamma + j)}\, p_0 \qquad n > c. \tag{5.6}$$

The constant p_0 is found by the normalization condition $\sum_{n=0}^{\infty} p_n = 1$. The queue length distribution q_n is given by

$$q_0 = \sum_{n=0}^{c} p_n = \frac{e^{\delta/\gamma}}{c!} \Gamma(c+1, \delta/\gamma)\, p_0 \tag{5.7}$$

$$q_n = p_{n+c} = \frac{\delta^{n+c}}{c!\gamma^c \prod_{j=1}^{n}(c\gamma + j)}\, p_0 \qquad n > 0, \tag{5.8}$$

where $\Gamma(z, a)$ is the upper incomplete Gamma function

$$\Gamma(z, a) = \int_a^{\infty} t^{z-1}\, e^{-t}\, dt,$$

which for integer values of $z = n$ satisfies

$$\Gamma(n+1, a) = n! e^{-a} \sum_{j=0}^{n} \frac{a^j}{j!}.$$

5.2 Wait Time Distribution

Let us now define the following events

$$\begin{aligned} W &= \text{Waiting in queue,} \\ A &= \text{Acquiring service.} \end{aligned}$$

An arriving unit has to wait in the queue if there are at least c units already in the system. Upon simplification, we obtain the probability that a unit waits in the queue as

$$P(W) = \sum_{n=c}^{\infty} p_n = \frac{e^{\delta}\, p_0}{\gamma^{c-1} \delta^{c(\gamma-1)} (c-1)!} \Gamma_l(c\gamma, \delta)\,, \tag{5.9}$$

where $\Gamma_l(a, z)$ is the lower incomplete Gamma function

$$\Gamma_l(z, a) = \int_0^a t^{z-1} e^{-t}\, dt.$$

The lower and upper incomplete Gamma functions are related to each other through the Gamma function:

$$\Gamma(z) = \Gamma_l(z, a) + \Gamma(a, z) = \int_0^\infty t^{z-1} e^{-t}\, dt.$$

We now compute $P(A, W)$, the probability that both events A and W occur, or in other words, the probability that a unit waits in the queue and acquires service without reneging. To do so, we first compute β_n, the probability that an arriving unit, upon finding $n \geq c$ units already in the system, receives service ($\beta_n = 1$ for $n < c$). Note that after such a unit, call it X, joins the queue, there are $n+1$ units in the system. Let us define an event as the departure of any unit in front of and including X (any future arrival does not affect the waiting time of X, and so is ignored in this explanation for simplicity). The time to the next event, whether it is service completion by one of the c units in service or reneging by one of the $n - c + 1$ units in queue, is exponentially distributed with parameter $c\mu + (n - c + 1)\alpha$. Thus, the probability that the next event is not the reneging of X is equal to the probability that either one of the units in service completes service, or one of the $n - c$ waiting units in front of X reneges, which is $[c\mu + (n - c)\alpha]/[c\mu + (n - c + 1)\alpha]$. Given that this event has occurred, there are now n units in the system, and so the probability that the next event will not be a reneging of X is β_{n-1}. By applying this processes repeatedly, and noting that each of the events is independent of the ones before due to the memoryless property of the exponential distribution, we can obtain the probability that X does not renege by iterating and noting the cancellations from successive terms ($n \geq c$):

$$\begin{aligned}
\beta_n &= \frac{c\mu + (n-c)\alpha}{\mu + (n-c+1)\alpha}\beta_{n-1} \\
&\vdots \\
&= \frac{c\mu}{c\mu + (n-c+1)\alpha}\beta_{c-1} \\
&= \frac{c\gamma}{c\gamma + n - c + 1}. \qquad (5.10)
\end{aligned}$$

Thus, by conditioning on the number of units already in the system at arrival, we can calculate the probability that a unit waits in the queue and successfully receives service:

$$\begin{aligned} P(A,W) &= \sum_{n=c}^{\infty} p_n \beta_n \\ &= \sum_{n=0}^{\infty} \frac{\delta^{n+c}}{c!\gamma^c \prod_{j=1}^{n}(c\gamma + j)}\, p_0 \cdot \frac{c\gamma}{c\gamma + n + 1} \\ &= \frac{\delta^c p_0}{(c-1)!\gamma^{c-1}} \sum_{n=0}^{\infty} \frac{\delta^n}{\prod_{j=1}^{n+1}(c\gamma + j)} \\ &= \Gamma_l(c\gamma + 1, \delta)\, \frac{\delta^{c-c\gamma-1} e^{\delta}}{(c-1)!\gamma^{c-1}}\, p_0. \end{aligned} \tag{5.11}$$

We now consider those units that join the queue and have a positive waiting time. Define

$$T_a = \text{time spent in queue by a unit acquiring service} \tag{5.12}$$

$$T_q = \text{time spent in queue by any unit that joins the queue.} \tag{5.13}$$

Note that T_q includes both those units that renege and those that acquire service. From these definitions it can be seen that

$$\begin{aligned} P\{t \le T_a \le t + dt\} &= P\{t \le T_q \le t + dt \,|\, (A,W)\} \\ &= P\{t \le T_q \le t + dt\,,\, (A,W)\}/P(A,W). \end{aligned} \tag{5.14}$$

Now, to compute the numerator on the right hand side of the above identity, consider a unit X that upon arrival finds j units already in the queue, an event that occurs with probability p_{c+j}. In this case, the probability that X survives to be served is β_{c+j} and the pdf of its total wait time in the queue is $(f_{j+1} * f_j * \cdots f_1)(t)$, where $f_k(t)$ is the pdf of the time until the next departure of any of the units in front of and including X, assuming X is the kth unit in the queue. Every one of these departures reduces the queue size by one, so that the pdf of the time to reach the server is the convolution of $f_k(t)$, for $k = j+1, \ldots, 1$. With k units in the queue, $f_k(t)$ is exponential with parameter $c\mu + k\alpha$:

$$\begin{aligned} f_k(t) &= (c\mu + k\alpha) e^{-(c\mu + k\alpha)t} \\ &= \alpha(c\gamma + k) e^{-(c\gamma + k)\alpha t}. \end{aligned} \tag{5.15}$$

To obtain the pdf $g_a(t)$ of the random variable T_a, that is, the wait time distribution of any unit in the queue that acquires service, we need to consider all the possibilities for j, conditioning on the number of units in the queue upon arrival:

$$g_a(t) = \frac{1}{P(A,W)} \sum_{j=0}^{\infty} p_{c+j}\beta_{c+j}(f_{j+1} * f_j * \cdots f_1)(t). \tag{5.16}$$

To compute the convolutions, we first transform the above expression. Let $g_a^*(s)$ and $f_j^*(s)$ be the Laplace Transform (LT) of $g_a(t)$ and $f_j(t)$, respectively. We can easily find that

$$f_k(s) = \frac{c\mu + k\alpha}{c\mu + k\alpha + s} = \frac{c\gamma + k}{c\gamma + k + s/\alpha}. \tag{5.17}$$

Then, using the fact that the LT of the convolution of two functions, $f_i(t)$ and $f_j(t)$, is given by the product of the LTs of the individual functions, $f_i^*(s)$ and $f_j^*(s)$, we obtain

$$\begin{aligned}
g_a^*(s) &= \frac{1}{P(A,W)} \sum_{j=0}^{\infty} p_{c+j}\beta_{c+j} \prod_{k=1}^{j+1} f_k^*(s) \\
&= \frac{1}{P(A,W)} \sum_{j=0}^{\infty} \frac{\delta^{c+j}}{c!\gamma_c \prod_{k=1}^{j}(c\gamma + k)} p_0 \cdot \frac{c\gamma}{c\gamma + j + 1} \cdot \prod_{k=1}^{j+1} \frac{c\gamma + k}{c\gamma + k + s/\alpha} \\
&= \frac{1}{P(A,W)} \frac{\delta^c}{(c-1)!\gamma^{c-1}} p_0 \sum_{j=0}^{\infty} \delta^j \prod_{k=1}^{j+1} \frac{1}{c\gamma + k + s/\alpha}.
\end{aligned}$$

Transforming the product into summation using partial fractions yields

$$g_a^*(s) = \frac{1}{P(A,W)} \frac{\delta^c}{(c-1)!\gamma^{c-1}} p_0 \sum_{j=0}^{\infty} \delta^j \sum_{k=1}^{j+1} \frac{(-1)^{k-1}}{(k-1)!(j+1-k)!} \cdot \frac{1}{c\gamma + k + s/\alpha}. \tag{5.18}$$

Using the linearity of LT, we can easily invert the last expression to get

$$\begin{aligned}
g_a(s) &= \frac{1}{P(A,W)} \frac{\delta^c}{(c-1)!\gamma^{c-1}} p_0 \sum_{j=0}^{\infty} \delta^j \sum_{k=1}^{j+1} \frac{(-1)^{k-1}}{(k-1)!(j+1-k)!} \alpha e^{-(c\gamma+k)\alpha t} \\
&= \frac{1}{P(A,W)} \frac{\alpha\delta^c}{(c-1)!\gamma^{c-1}} p_0 e^{-(c\gamma+1)\alpha t} \sum_{j=0}^{\infty} \delta^j \sum_{k=0}^{j} \frac{(-1)^k}{k!(j-k)!} \left(e^{-\alpha t}\right)^k,
\end{aligned}$$

which upon using the binomial theorem yields

$$\begin{aligned} g_a(s) &= \frac{1}{P(A,W)} \frac{\alpha\delta^c}{(c-1)!\gamma^{c-1}}\, p_0 e^{-(c\gamma+1)\alpha t} \sum_{j=0}^{\infty} \frac{\delta^j}{j!} \sum_{k=0}^{j} \left(\begin{array}{c} j \\ k \end{array} \right) \left(-e^{-\alpha t}\right)^k \\ &= \frac{1}{P(A,W)} \frac{\alpha\delta^c}{(c-1)!\gamma^{c-1}}\, p_0 e^{-(c\gamma+1)\alpha t} \sum_{j=0}^{\infty} \frac{\delta^j}{j!} \left(1-e^{-\alpha t}\right)^j \\ &= \frac{1}{P(A,W)} \frac{\alpha\delta^c}{(c-1)!\gamma^{c-1}}\, p_0 e^{-(c\gamma+1)\alpha t}\, e^{\delta(1-e^{-\alpha t})}. \qquad (5.19) \end{aligned}$$

Using equations (5.6) and (5.11) to substitute for p_0 and $P(A,W)$, respectively, we obtain

$$g_a(t) = \frac{\alpha\delta^{c\gamma+1}}{\Gamma_l(c\gamma+1,\delta)} \exp\{-(c\gamma+1)\alpha t - \delta e^{-\alpha t}\}. \qquad (5.20)$$

By defining the rescaled parameter $\tilde{\gamma} = c\gamma = c\mu/\alpha$, expression (5.20) can be written as

$$g_a(t) = \frac{\alpha\delta^{\tilde{\gamma}+1}}{\Gamma_l(\tilde{\gamma}+1,\delta)} \exp\{-(\tilde{\gamma}+1)\alpha t - \delta e^{-\alpha t}\}. \qquad (5.21)$$

It can be verified that $g_a(t)$ is normalized. For the special case of $c = 1$, that is, an $M/M/1$ queue with reneging, we have [30]

$$g_a(t) = \frac{\alpha\delta^{\gamma+1}}{\Gamma_l(\gamma+1,\delta)} \exp\{-(\gamma+1)\alpha t - \delta e^{-\alpha t}\}, \qquad (5.22)$$

which is also a special case (see the Appendix) of the result of Ancker and Gafarian [3] for the $M/M/1$ queue with balking and reneging. Note that the distribution for the wait time of *any* unit acquiring service is given by

$$h_a(t) = c_0\delta(t) + (1-c_0)g_a(t), \qquad (5.23)$$

where $c_0 = p_0 + \cdots + p_{c-1}$ is the free-server probability.

The mean wait time in the queue for those units acquiring service can be obtained by computing the first moment via $-\frac{d}{ds}g_a^*(s)\big|_{s=0}$:

$$E\{W_a\} = \frac{1}{P(A,W)} \frac{\delta^c}{(c-1)!\gamma^{c-1}}\, p_0 \sum_{j=0}^{\infty} \delta^j \sum_{k=0}^{j} \frac{(-1)^k}{k!(j-k)!} \cdot \frac{\alpha^{-1}}{(c\gamma+k+1)^2},$$

which upon interchanging the order of summations yields

$$\begin{aligned}
E\{W_a\} &= \frac{\alpha^{-1}}{P(A,W)}\frac{\delta^c}{(c-1)!\gamma^{c-1}}\,p_0\sum_{k=0}^{\infty}\frac{(-1)^k}{k!(c\gamma+k+1)^2}\sum_{j=k}^{\infty}\frac{\delta^j}{(j-k)!} \\
&= \frac{\alpha^{-1}}{P(A,W)}\frac{\delta^c}{(c-1)!\gamma^{c-1}}\,p_0\sum_{k=0}^{\infty}\frac{(-\delta)^k}{k!(c\gamma+k+1)^2}\sum_{j=0}^{\infty}\frac{\delta^j}{j!} \\
&= \frac{\alpha^{-1}}{P(A,W)}\frac{\delta^c e^{\delta}}{(c-1)!\gamma^{c-1}}\,p_0\sum_{k=0}^{\infty}\frac{(-\delta)^k}{k!(c\gamma+k+1)^2} \\
&= \frac{\delta^{\tilde{\gamma}+1}\alpha^{-1}}{\Gamma_l(\tilde{\gamma}+1,\delta)}\sum_{k=0}^{\infty}\frac{(-\delta)^k}{k!(\tilde{\gamma}+k+1)^2}\,. \qquad (5.24)
\end{aligned}$$

So far we have assumed that the reneging parameter is known in advance. However, since the database of B.C. acute care hospitals can only provide us with the percentage of reneged patients, we cannot directly obtain the reneging parameter. We next derive a relationship between the reneging parameter α and the percentage of reneged patients κ which allows us to resolve this issue.

5.3 The Reneging Parameter

Let κ be the probability that an arrival reneges. From the definition of β_n, given that there are n units in the system upon arrival, the probability that a unit reneges is given by $1-\beta_n$. Hence, by conditioning on the number of units in the system at arrival and using equation (5.10) we obtain

$$\begin{aligned}
\kappa &= \sum_{n=c}^{\infty}(1-\beta_n)\,p_n \\
&= \sum_{n=c}^{\infty}\frac{n-c+1}{c\gamma+n-c+1}\,p_n.
\end{aligned}$$

Using equation (5.2) to replace p_n by p_{n+1}, we can write the previous expression as

$$\kappa = \frac{\alpha}{\lambda}\sum_{n=c}^{\infty}(n-c+1)\,p_{n+1},$$

which can be written in terms of the mean queue length L:

$$\begin{aligned}\kappa &= \frac{\alpha}{\lambda}\sum_{n=0}^{\infty} n q_n \\ &= \frac{\alpha}{\lambda} L. \qquad (5.25)\end{aligned}$$

This is a nonlinear equation in α, due to the dependence of L on α. To compute L we may use equations (5.7) and (5.8) to compute q_n. However, this approach poses a challenge for large values for c and λ, since in such cases the term $\frac{1}{c!}\Gamma(c+1, \delta/\gamma)$ in q_0 involves dividing two very large numbers (using numerical software, such as MATLAB, this can result in `Inf/Inf`, which returns `NaN` – Not a Number). To overcome this obstacle, we use an approximation motivated by Stirling's formula, which gives the asymptotic behaviour of $n!$ for large values of n:

$$\Gamma(n+1) = n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n . \qquad (5.26)$$

Starting from (5.7), using the series representation of the incomplete Gamma function and making use of Stirling's formula, we obtain the following approximation to q_0:

$$\begin{aligned} q_0 &= \frac{e^{\delta/\gamma}}{c!}\Gamma(c+1, \delta/\gamma)\, p_0 \\ &= \sum_{k=0}^{c} \frac{(\delta/\gamma)^k}{k!}\, p_0 \\ &\sim \sum_{k=0}^{r-1} \frac{(\delta/\gamma)^k}{k!}\, p_0 + \sum_{k=r}^{c} \frac{(e\delta/k\gamma)^k}{\sqrt{2\pi k}}\, p_0 \, , \qquad (5.27)\end{aligned}$$

where r is chosen large enough that Stirling's formula is a good approximation, but not so large as to introduce numerical errors. A good choice would be $r = 10$, which gives the relative error of

$$\frac{1}{r!}\left| r! - \sqrt{2\pi r}\left(\frac{r}{e}\right)^r \right| = 0.0083 \, .$$

In the above formula we see that for large values of k the expressions $(\delta/\gamma)^k/k!$ and $(e\delta/k\gamma)^k/\sqrt{2\pi k}$ are asymptotically equal. However, when δ/γ and c are large, computationally it is more accurate to calculate the second expression, which avoids division of very large numbers. Note that since $\delta/\gamma = \lambda/\mu$, this approximation is applicable to queues with many servers (large c) operating under heavy load (large λ/μ).

In a similar manner, by rearranging terms and applying Stirling's formula to $c!$, we obtain an approximation to q_n for $n > 0$ which is numerically stable:

$$\begin{aligned} q_n &= \frac{\delta^{n+c}}{c!\gamma^c \prod_{j=1}^{n}(c\gamma + j)}\, p_0 \\ &= \frac{\left(\frac{\delta}{\gamma}\right)^c \left(\frac{\delta}{c\gamma}\right)^n}{c! \prod_{j=1}^{n}\left(1 + \frac{j}{c\gamma}\right)}\, p_0 \\ &\sim \frac{e^c \left(\frac{\delta}{c\gamma}\right)^{n+c}}{\sqrt{2\pi c} \prod_{j=1}^{n}\left(1 + \frac{j}{c\gamma}\right)}\, p_0. \end{aligned} \tag{5.28}$$

Note that p_0 can simply be computed by enforcing the normalization condition. We then used MATLAB's `fzero` function to calculate α from equation (5.25).

5.4 Reneging in Service

Here we consider the possibility of units' leaving the service station before their service is completed. We assume that the in-service reneging time is exponentially distributed, with parameter α'. This greatly simplifies the analysis, for now the length of stay in the service station is exponential with mean $W' = (\alpha' + \mu)^{-1}$, and only a fraction $\mu/(\alpha' + \mu)$ actually complete service, while the rest renege. In other words, the probability of reneging inside the service station is given by

$$\kappa' = \alpha' W' \,. \tag{5.29}$$

We note that equation (5.25) can also analogously be written as

$$\kappa = \alpha W \tag{5.30}$$

using Little's Theorem, where W denotes the mean wait time in the queue by all units, whether they renege or not.

Chapter 6

Two Queues in Tandem

In our model of the hospital, the ICU and the MU have a finite number of beds and no waiting space in between. In other words, a patient ready to leave the ICU would not be able to if the MU is full. This causes blocking of the patient in the ICU. If an infinite queue were allowed in between the two units and if the reneging process were ignored, this queueing network would be the same as that of section 2.8, resulting in a product form solution for the equilibrium distribution of each station, independent of the other.

Product form solutions of queueing networks are very important from a computational standpoint, for they provide a way to decompose an otherwise very large state space into independent subspaces. Even for a small number of service stations and a moderately small number of servers in each, storing the whole state space and performing computations on it becomes a daunting task at best. However, if the stations can be considered as independent, allowing for a product form solution, we can perform the computation on the state space of each of the individual stations independently of the others, resulting in a great reduction in the computational complexity. Hence, much work has been devoted to approximating non-product-form queueing networks using product-form networks.

In most studies, the individual stations are treated as independent queues but with modified arrival or service rates. Perros [27] and Balsamo *et al.* [4] have collected a vast literature on the approximation of queueing networks with blocking, but they consider only single-server stations. There are other decomposition schemes with single-server assumptions that are not considered in these two books. Among them are the work by Boxma *et al.* [7], where they approximate each single-server node by a superposition of two $M/M/1/N$ queue distributions, each of which is valid in a specific phase ($N-1$ is the maximum waiting space for each station). The probability of being in each phase and the parameters of each of the $M/M/1/N$ queues are determined by solving a set of nonlinear equations in an iterative fashion. In fact, most of the research in this area involves solving a set of nonlinear equations for the parameters of the service stations of the decomposed network, since the blocking phenomenon causes interaction between them.

The literature on multi-server finite capacity queueing network is more sparse. In section 6.6.2 of [6], Bose describes an algorithm, called the Maximum Entropy Method, for the approximate decomposition of a queueing network with finite capacity service stations. The Expansion Method is another approximation algorithm, which though originally proposed for single-server queueing networks, was extended by Jain *et al.* [19] to multi-server stations. In this method, the original network is expanded by inserting an $M/M/\infty$ queue between every two adjacent stations; any blocked unit is served in this intermediate node and retries for entry into the next station after its service is completed. Both of these methods, like the previous ones, rely on solving a set of nonlinear equations iteratively to find the desired parameters. The power of such algorithms is best realized in networks of considerably large size, so that the computational advantage of the iterative schemes surpasses that of numerically solving the whole network.

In this project, having a network of only two stations, we focus on simpler approaches, which do not rely on iterative schemes. In [20], Koizumi *et al.* present an approximation method, which decomposes the network by computing effective service rates for the blocked stations, and an effective arrival rate by considering the flow from neighbouring stations. The effective service rate is obtained by incorporating the average wait time in the upstream stations into the actual service rate of that station. In another method, by Korporaal *et al.* [21], the authors use the average queue length from upstream stations to compute the mean number of blocked servers, which when subtracted from the total number of servers, yields the average number of available servers. Both of these methods are discussed at length

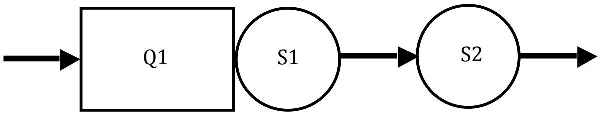

Figure 6.1: Two queues in tandem

in section 6.2. The advantages of these methods are ease of implementation and intuitive understanding.

The queueing system considered in this thesis is composed of only two finite capacity stations placed in tandem (Fig. 6.1). Consider the situation in which customers, after being served at the first station (S_1), must proceed to the next (S_2) for further service. If all servers are occupied in S_2, these customers cannot proceed and get blocked, staying at S_1 and occupying their servers. Upon completing service in S_2, customers leave the system, in which case the earliest blocked customer in S_1, if any, proceeds to S_2. We assume that inter-arrival and service times are all exponential and that blocking occurs after service; note that in certain manufacturing problems, blocking before service is used, in which if a unit arriving at a service station finds an upstream station full, then it does not begin its service and is blocked.

The following quantities completely specify this tandem queueing system:

- λ: arrival rate into S_1
- μ_i: service rate in S_i for $i = 1, 2$
- c_i: number of servers in S_i for $i = 1, 2$

Our goal is to solve for the steady-state distribution of the number of units waiting in the queue and also for the distribution of the wait time. We first present a numerical solution for this queueing system that is exact to machine precision, and then present approximate analytical solutions.

6.1 Numerical Solution

Consider the extended state space (m, n) for $m = 0, 1, 2, \ldots$ and $n = 0, 1, 2, \ldots, c_1 + c_2$. Here, the variable m denotes the number of units in the S_1 subsystem, that is, those waiting in the queue before S_1 in addition to those being served in the station. The variable n denotes the number of units in the S_2 subsystem, where in this case units waiting for this station are those that are blocked in S_1. In other words, n is the sum of the number of units being served in S_2 and the number of units blocked in S_1. As a result, values of $n \leq c_2$ represent the number of busy servers in S_2, while for $n > c_2$, all the c_2 servers in S_2 are busy and $n - c_2$ represents the number of units blocked in S_1. By defining

$$r = \min(m, c_1) \tag{6.1}$$

as the number of occupied (but not necessarily busy) servers in S_1, we can write the state transitions and their associated rates as follows:

Table 6.1: Transition rates of tandem queue

state transition	rate	condition
$(m, n) \to (m + 1, n)$	λ	$m \geq 0\,,\, n \geq 0$
$(m, n) \to (m - 1, n + 1)$	$r\mu_1$	$m > 0\,,\, n < c_2$
$(m, n) \to (m, n + 1)$	$(r + c_2 - n)\mu_1$	$m > 0\,,\, n \geq c_2$
$(m, n) \to (m, n - 1)$	$n\mu_2$	$m \geq 0\,,\, 0 \leq n \leq c_2$
$(m, n) \to (m - 1, n - 1)$	$c_2\mu_2$	$m > 0\,,\, n > c_2$

The transition $(m, n) \to (m + 1, n)$ is due to the arrivals into the system. The second transition occurs when a unit completes its service in S_1 and goes to S_2, which is possible only when there are free servers there, i.e. $n < c_2$. If there are no free servers in S_2, then the unit cannot leave S_1 and gets blocked, and so the n parameter is incremented to indicate an increase in the number of blocked units, while m remains the same. The number of busy servers, excluding those which are occupied by blocked units, is given by $r - (n - c_2)$, which accounts for the indicated rate of change. Now, if a unit departs from S_2 when there are no blocked units in S_1, i.e. $0 \leq n \leq c_2$, then state (m, n) changes to $(m, n-1)$, which happens at rate $n\mu_2$, as there are n busy servers in S_2. However, if there are blocked units in S_1, then m also decreases, as one of the blocked units moves to S_2. In this case, although S_2 still remains at full capacity, the counter n decreases by one to indicate a decrease in the number of blocked units, leading to state transition $(m, n) \to (m-1, n-1)$.

Using these rates of change, we construct the rate matrix Q and obtain the joint stationary distribution $\tilde{p}_{mn}$ (see section 2.1), from which the stationary distribution of the *actual* number of units in each of the subsystems, S_1 and S_2, can be constructed by disregarding the blocked-units counter embedded in the second parameter space:

$$\begin{aligned} p_{mn} &= \tilde{p}_{mn}\,, \qquad\qquad n < c_2\,, \\ p_{mc_2} &= \sum_{n=c_2+1}^{c_1+c_2} \tilde{p}_{mn}. \end{aligned}$$

To obtain the distribution $w(t)$ of the wait time in the queue (in front of S_1), we use (2.17). From p_{mn} we can obtain the queue length distribution q_j as follows:

$$q_0 = \sum_{m=0}^{c_1} \sum_{n=0}^{c_2} p_{mn}, \tag{6.2}$$

$$q_j = \sum_{n=0}^{c_2} p_{c_1+j,n} \qquad \text{for } j > 0, \tag{6.3}$$

where we used $j = \max(m - c_1, 0)$. Now, following the discussion in section 2.6, we can find $w(t)$ numerically. We demonstrate this using an example. Let us choose

$$\begin{aligned} \lambda &= 2, \\ c_1 &= 10, \quad c_2 = 20, \\ \mu_1 &= \tfrac{1}{3}, \quad \mu_2 = \tfrac{1}{7}, \end{aligned}$$

from which we numerically compute the PGF $P(z) = \sum_{m=0}^{\infty} q_j z^j$ using $z = n\Delta z$ for $\Delta z = 0.01$ and $n = -100, -99, \ldots, 99, 100$. Then, by using the `ratpolyfit` function with polynomial degree $k = 1$ for both the numerator and denominator we arrive at the rational approximation

$$\tilde{P}_1(z) = \frac{0.567\,z - 0.921}{0.646\,z - 1}. \tag{6.4}$$

The error in the approximation in this case is

$$e_0 = ||P(z) - \tilde{P}_1(z)||_\infty = 5.1 \times 10^{-4}. \tag{6.5}$$

Using $w^*(s) = P(1 - s/\lambda)$, the approximate LT of the wait time distribution with $k = 1$ is given by

$$\tilde{w}_1^*(s) = \frac{1 + 0.800\,s}{1 + 0.912\,s} \tag{6.6}$$

Now, by inverting the LT we obtain

$$\tilde{w}_1(t) = 0.878\,\delta(t) + 0.134e^{-1.097t} \tag{6.7}$$

as an approximation to the pdf of the wait time distribution. The error estimate e_1, obtained by comparing the coefficient of the delta function to the probability that a unit upon arrival at S_1 finds a free server (see section 2.6), is given by

$$e_1 = \left| \sum_{m=0}^{c_1-1} q_m - 0.878 \right| = 6.95 \times 10^{-4}. \tag{6.8}$$

In addition, the measure of how well our estimated pdf is normalized is given by

$$e_2 = \left| 1 - \int_0^\infty \tilde{w}_1(t)\,dt \right| = 5.1 \times 10^{-4}. \tag{6.9}$$

In a similar fashion, we can obtain the wait time distributions for other polynomial degrees k. Below we lists the expressions $\tilde{w}_k(t)$ obtained for $k = 1, 2, 3, 4$

$$\begin{aligned}
k = 1 : \tilde{w}_1(t) &= 0.878\delta(t) + 0.130e^{-1.075t}, \\
k = 2 : \tilde{w}_2(t) &= 0.877\delta(t) + 0.047e^{-.839t} + 0.092e^{-1.376t}, \\
k = 3 : \tilde{w}_3(t) &= 0.877\delta(t) + 0.083e^{-1.393t} + 0.040e^{-.930t} + 0.016e^{-.806t}, \\
k = 4 : \tilde{w}_4(t) &= 0.877\delta(t) + 0.005e^{-1.700t} + 0.090e^{-1.347t} + 0.045e^{-.832t} + 0.115e^{4.0576t}.
\end{aligned}$$

As can be seen, for $k = 4$, the approximated wait time pdf grows exponentially, which is unacceptable. In fact, the same behaviour repeats for values of $k \geq 4$. This is in line with our discussion in section 2.6 where we stated that values of k close to 1 should be chosen, since with larger k the conditioning of the rational approximation becomes worse. Because the error in the rational approximation translates to an error in the roots of the polynomials, and since the exponents in the LT inversion are the roots of the polynomial in the denominator, exponentially growing terms may emerge as a result of a poor approximation.

Ignoring the $k = 4$ case, the error estimates corresponding to $k = 2$ and $k = 3$ are

$$k = 2: \quad e_0 = 5.0 \times 10^{-6}, \quad e_1 = 6.7 \times 10^{-6}, \quad e_2 = 4.1 \times 10^{-2},$$

$$k = 3: \quad e_0 = 5.3 \times 10^{-4}, \quad e_1 = 3.6 \times 10^{-5}, \quad e_2 = 5.3 \times 10^{-4}.$$

Comparing these to the $k = 1$ case, we may conclude that both the $k = 2$ and $k = 3$ cases give very good results. Furthermore, taking $\tilde{w}_3$ as the best approximation, the following relative errors indicate that all three approximations are quite similar numerically, which may not be evident from the mathematical expressions themselves:

$$\begin{aligned} \frac{||\tilde{w}_3 - \tilde{w}_1||_2}{||\tilde{w}_3||_2} &= 6.0 \times 10^{-4} \\ \frac{||\tilde{w}_3 - \tilde{w}_2||_2}{||\tilde{w}_3||_2} &= 1.3 \times 10^{-4} \end{aligned}$$

It can thus be seen that in using the method of section 2.6 for the purpose of finding the wait time pdf from the queue length distribution of an $M/G/c$ queue, one can start with a $k = 1$ approximation, measuring the error estimates e_0, e_1, and e_2, and repeat the process for incrementally larger values of k until the inverted LT yields exponentially growing terms, at which time we can stop the process and choose the best of the approximations using the error estimates e_1, e_2 and e_3.

6.2 Approximate Solution

In this section, we review the work by Korporaal *et al.* [21], where the authors use a queueing network model to predict the probability that a criminal has to be sent home because of a shortage of cells in Dutch prisons. They model the penitentiary system in the Netherlands, consisting of several prisons or institutions of different types connected to each other, using a network of finite capacity queues with blocking. Each prison (service station) has a finite number of cells (servers). We discuss their approximation method in the context of two queues in tandem.

In their algorithm, the two stations, S_i, for $i = 1, 2$, are approximated by two queues of type $M(\lambda)/M(\mu_i)/s_i/N_i$, where the symbol $M(\cdot)$ represents a Poisson process with the argument as the parameter. The parameters are chosen as follows: Since no unit is lost in going from one station to the other, the inflow rate of units must be the same as the outflow at equilibrium. Thus, the same parameter λ is chosen as the arrival rate to both stations[1]. In our case, since an infinite number of patients are allowed to queue at the first station, reflecting the nature of the ER, we have that $N_1 = \infty$. The buffer size for the second station is set to the number of servers in that station plus those of the previous station, since if all the server in S_1 are blocked, then there are up to c_1 units waiting to enter S_2 (this is analogous to how the state space was chosen in the previous section). Thus, $N_2 = c_1 + c_2$. The effective number of servers in station S_i is taken to be s_i to reflect the fact that some of the available c_i servers, due to blocking, cannot actually serve a customer. Thus, s_i at a station with blocking is smaller than the actual number of servers c_i. To incorporate this fact into their model, the authors suggest using $s_1 = c_1 - Q_2$, where Q_2 is the mean queue length at S_2. The reason for this choice is that the queue for S_2 is formed by the blocked units in S_1, so the mean queue length at S_2 represents the mean number of blocked servers in S_1. To deal with the restriction that s_i must be an integer, and the fact that Q_2 may have non-integer values, it was suggested to use a linear combination of two queue distributions, one with $\lceil s_1 \rceil$ servers and the other with $\lfloor s_1 \rfloor$ servers (as the authors do not mention, we assume that the weights are $s_1 - \lfloor s_1 \rfloor$ and $\lceil s_1 \rceil - s_1$, respectively). At S_2, since no blocking exists, $s_2 = c_2$.

One of the special aspects of the penitentiary system modelled in [21] is that the scheduled term of imprisonment in a prison is diminished by the time of waiting for a transfer (waiting time in the queue) when the prison turns out to be full at the planned time of the transfer. So, an adjustment is made to the service rates at each station after the first, based on the units' waiting time in the previous station. In our model, however, we have assumed that the service times at each station are independent of what happened before. Thus, μ_i are the actual service rates at each station.

Having defined all the parameters, the approximate queue length distributions $p_i(n)$ for S_i

[1] In this section, we ignore the inflow of external units into S_2. The addition of this stream of units corresponding to the non-ICU cases arriving to the ED is considered in chapter 7.

can be computed from the analytical formulas for $M(\lambda)/M(\mu_i)/s_i/N_i$. Note that we must first compute $p_2(n)$, whose parameters are independent of the first station. Then by computing the mean queue length Q_2 we can determine the effective number of servers s_1 in the first station.

It is important to understand that the approximation in the above algorithm comes about as a result of two main assumptions. Firstly, it is assumed that the blocking phenomenon can be accounted for solely by decreasing the total number of servers in a station experiencing blocking (S_1). This reduction is given by the mean queue length at the downstream station (S_2), since the waiting units at S_2 are viewed as the blocked ones in S_1. However, since in general the mean queue length is not integer valued, resulting in a non-integer number of servers, it is secondly assumed that this inconsistency can be fully accounted for by taking a linear combination of two queue distributions with integer valued server counts.

Considering the second assumption, we now suggest a modification to this method that improves the approximation. Instead of taking a linear combination of two queue distributions, we note that the quantities μ_1 and c_1 often occur in the form $c_1\mu_1$ in analysis of queueing system (this quantity is the maximum rate of departure that can be achieved when all c_1 servers are busy). Thus, instead of modifying $c_1 \cdot \mu_1$ to $s_1 \cdot \mu_1$, which implies changing the number of servers with the service rate fixed, we suggest the alternative $c_1 \cdot \left(\frac{s_1}{c_1}\mu_1\right)$; in other words, we let the number of servers remain the same, but modify the service rate to $\frac{s_1}{c_1}\mu_1$. Below we make a comparison of this new approach to that of Korporaal *et al.* [21].

Having an exact numerical solution to this queueing system, we can compare the accuracy of the two methods. Let $p_1^{(i)}$ and $Q_1^{(i)}$ be the distribution and the mean of the number of units in S_1 using Method i as described below:

- Method 1: numerical solution
- Method 2: approximation of Korporaal *et al.* using μ_1 as service rate and interpolation of two queue distributions with $\lfloor s_1 \rfloor$ and $\lceil s_1 \rceil$ servers
- Method 3: approximation using $\frac{s_1}{c_1}\mu_1$ as the service rate and c_1 servers

	Method 1		Method 2		Method 3	
c_2	$Q_1^{(1)}$	p_b	$Q_1^{(2)}$ rel. err.	$\lvert p_1^{(2)} - p_1^{(1)} \rvert_1$	$Q_1^{(3)}$ rel. err.	$\lvert p_1^{(3)} - p_1^{(1)} \rvert_1$
15	6.970	70.4%	79.1%	61.1%	76.4%	60.3%
16	2.492	50.9%	80.7%	39.1%	74.4%	35.9%
17	1.049	35.2%	74.3%	24.3%	66.0%	20.6%
18	0.532	23.6%	58.9%	13.3%	53.2%	11.8%
19	0.323	15.4%	41.1%	7.3%	37.7%	6.5%
20	0.231	9.7%	25.1%	4.0%	23.3%	3.5%
21	0.189	5.9%	13.6%	2.1%	12.7%	1.9%
22	0.169	3.4%	6.8%	1.1%	6.3%	1.0%
23	0.160	2.0%	3.2%	0.6%	3.0%	0.5%
24	0.156	1.1%	1.4%	0.3%	1.3%	0.2%
25	0.154	0.6%	0.6%	0.1%	0.6%	0.1%

Table 6.2: Error in tandem queue approximation using Methods 2 and 3

For the runs, we kept all the quantities except c_2 constant:

$$\begin{array}{rclrcl} \lambda & = & 2, & & & \\ \mu_1 & = & \frac{1}{3}, & \mu_2 & = & \frac{1}{7}, \\ c_1 & = & 10, & c_2 & = & 15, \ldots, 25. \end{array}$$

By varying the number of servers in S_2 from $c_2 = 15$ to $c_1 = 25$ we guarantee that the tests incorporate both the scenarios in which blocking happens rarely and also those in which blocking is a dominant factor (for $c_2 < 15$ the system does not have an equilibrium). The blocking effect can be measured by the blocking probability

$$p_b = \sum_{n=c_2}^{c_1+c_2} p_2(n)\,. \tag{6.10}$$

We only list the values of p_b computed by the numerical solution. Note that the approximations for p_b are the same using method 2 or 3, as both methods treat S_2 in the same way, and only differ in computing the distribution for S_1. For each value of c_2, the table lists the relative error in approximations $Q^{(i)}$ for $i = 2, 3$. In addition, we compare the accuracy in the whole distribution using L_1-norm $\lvert p_1^{(I)} - p_1^{(1)} \rvert_1$ for $i = 2, 3$ – this norm is appropriate here since $\lvert p_1^{(i)} \rvert_1 = 1$. The results from table 6.2 show that the modification of the method

	Method 1		**Method 4**	
c_2	$Q_1^{(1)}$	p_b	$Q_1^{(2)}$ rel. err.	$\left\lvert p_1^{(2)} - p_1^{(1)} \right\rvert_1$
15	6.970	70.4%	57.1%	50.0%
16	2.492	50.9%	56.4%	35.0%
17	1.049	35.2%	49.3%	22.7%
18	0.532	23.6%	38.8%	13.9%
19	0.323	15.4%	26.5%	8.0%
20	0.231	9.7%	15.6%	4.5%
21	0.189	5.9%	7.9%	2.4%
22	0.169	3.4%	3.6%	1.3%
23	0.160	2.0%	1.5%	0.6%
24	0.156	1.1%	0.6%	0.3%
25	0.154	0.6%	0.2%	0.2%

Table 6.3: Error in tandem queue approximation using Method 4

of Korporaal *et al.* (Method 3) performs better than their original one (Method 2).

We now discuss another approximation algorithm due to Koizumi *et al.* [20], which was briefly addressed previously. In this method, we treat S_2 as before, i.e. as an $M/M/c_2/(c_1+c_2)$ queue, even though the authors in their decomposition algorithm assume an infinite queue allowed before each station; we found that the finite capacity restriction gives much more accurate results. However, our modification, enforcing a finite queueing capacity for S_2, cannot readily be applied to a general network in which there may be multiple service stations upstream from a given station, making it ambiguous as to how to choose the finite capacity. The only difference from Method 2 comes in computing the distribution of S_1. Koizumi *et al.* suggest using c_1 as the service rate, and modifying the service rate by incorporating the waiting time in the queue for S_2. Note that the units waiting in the queue for S_2 are actually those who are blocked in S_1. Thus, W_2, the mean waiting time in queue for S_2, is the mean blocking time in S_1, and so the effective mean service time in S_1 is approximated by

$$\frac{1}{\tilde{\mu}_1} = \frac{1}{\mu_1} + W_2 \,. \tag{6.11}$$

The quantity W_2 can easily be found by computing the mean queue length before S_2 from the approximate queue distribution and applying Little's theorem. Table 6.3 shows the results obtained from this algorithm, using the same parameter values as before, which we refer to as Method 4.

Comparing these results to those of Method 3, we observe that the technique of Koizumi *et al.* [20] computes better approximations to the mean queue length, but it performs worse in computing the whole distribution for values of $c_2 > 16$.

In the next chapter we apply the approximation techniques from this section to treat the ICU-MU interaction for the purpose of estimating the required number of beds in each unit.

Chapter 7

Bed Estimation

In this chapter, we first describe an abstract queueing system that represents the ED-ICU-MU network. Then, using analytical methods, we estimate the required number of beds in the ICU and the MU that would guarantee the following two TTAs for arriving emergency patients: 90% within 1 hour for those going to the ICU and 80% within 6 hours for those going to the MU. The analytical work also enables us to estimate some of the unknown parameters, which are then used in the simulation to obtain more accurate results. The discrete-event simulation precisely represents the queueing network described; however, since searching in the two dimensional parameter space of the number of beds in the ICU and MU is computationally very expensive using simulation, the estimates from the analytical work are used to provide a guess for the neighbourhood of the search.

7.1 Queueing Network Model

The model queueing network consists of two stations, S_1 and S_2 (see Figure 7.1). There are two classes of units in this system: for $i = 1, 2$ class i units are those who require service at S_i upon their arrival. If there are no servers available, the units wait in the queue, but while waiting, any unit may renege and leave the queue if it does not receive service within a certain time, called reneging time. In addition, any unit being served in S_1 has the potential to renege. If a unit successfully completes service in S_1, then it attempts to go to S_2. However, if all servers are busy there, the unit becomes blocked, and occupies its server

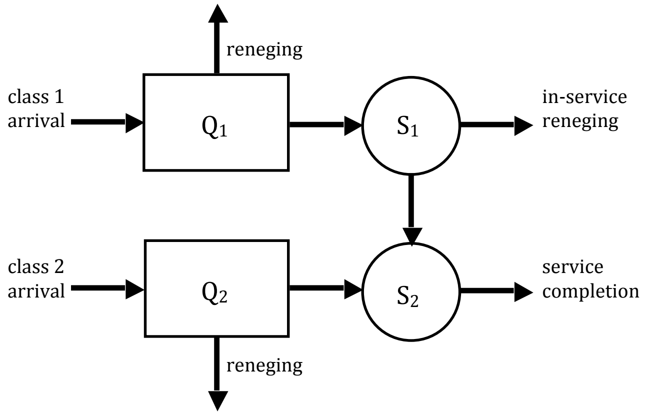

Figure 7.1: The tandem $S_1 - S_2$ queueing network

in S_1 while waiting to be admitted; note that we assume that even during the period of blocking a unit may renege from S_1 and leave the system. When a server becomes available in S_2, blocked units in S_1 are given priority access over class 2 units. On average, both classes of units inside S_2 are assumed to require the same amount of service, which when completed, results in their departure from the system.

As before, we assume that interarrival times, service times, and reneging times are all independent and exponentially distributed with constant rates. With these assumptions, the following parameters completely specify the system characteristics:

c_i : number of servers in S_i

λ_i : mean arrival rate to S_i

μ_i : mean service rate in S_i

α_i : mean reneging rate in Q_i

α_1' : mean reneging rate in S_1

We now make the connection between this abstract queueing network and the ED-ICU-MU model. The ICU and the MU are represented by S_1 and S_2, respectively, with beds being the servers. The emergency room is modelled as the combined queue leading to the two stations. The queue for S_1 represents the group of emergency patients in the ED waiting for a bed in ICU, and the reneging in this queue or S_1 itself corresponds to medical fatalities in the ED or the ICU, respectively. Similarly, the queue before S_2 represents the waiting

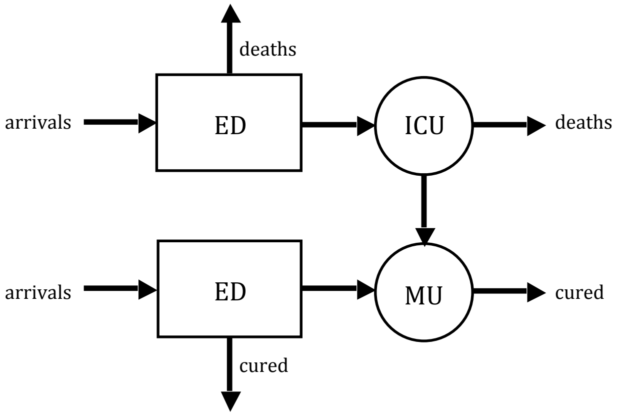

Figure 7.2: The network representation of the ED-ICU-MU

patients in the ED requiring a bed in the MU, and the reneging in this queue corresponds to treatment completions in the ED. The reason blocked class 1 units are given priority access to S_2 over class 2 units is that, as mentioned previously, we assume that doctors often try to free the ICU beds as soon as possible, both to keep the cost of the ICU bed maintenance low and also to allow fast access to the ICU.

We now proceed to determine the values of system parameters defined earlier from the database provided to us by the British Columbia (B.C.) Ministry of Health Services.

7.1.1 Parameter Specification

For the purpose of this project, we consider a hospital in B.C. of typical size (the name must remain undisclosed). From the data collected over the period of a year, the following quantities can be computed (for consistency, we use the notation and terminology used earlier in defining the abstract queueing network, instead of referring to the ICU or the MU):

N_i = number of units arrived at S_i in a year

W_i' = average length of stay in S_i

κ_i = percentage of units reneged from Q_i

κ_1' = percentage of units reneged in S_1

The data analysis performed by Vertesi [34] on the 2006-07 database provided us with the following estimated values:

$$\begin{aligned} N_1 &= 792, & N_2 &= 9979, \\ W_1' &= 4.44\,\text{days}, & W_2' &= 6.98\,\text{days}, \\ \kappa_1 &= 2.39\%, & \kappa_2 &= 2.32\%, \\ \kappa_1' &= 15.8\%. \end{aligned}$$

Note that for a queueing network in which blocking and reneging does not occur, the average length of stay in each station is the same as the inverse of the mean service rate for that station. However, due to blocking, the mean length of stay is larger by an amount approximately equal to the mean blocking time. This, in effect, is the essence of the method of Koizumi *et al.* [20] (Method 4) for decomposing tandem queues as discussed in chapter 6, characterized by the use of equation (6.11) to compute the effective service time. Since the data already provide the lengths of stays, that is the blocking time plus the treatment time, in using Method 4, the effective service rate is simply given by the inverse of the length of stay. However, for our simulation, the actual service rate needs to be determined. Moreover, due to the fact that the recorded lengths of stays in the ICU include those of the patients that died in this unit, the service rate in the ICU must be adjusted to account for this. These issues are addressed in the next section, where we analyze S_1 and S_2 in isolation.

7.2 Analytical Approximation

Here we use the decomposition method described in chapter 6, along with the reneging results of chapter 5, to obtain the approximate number of servers required in each of the stations S_1 and S_2 in order to guarantee the required target access rates. In trying to use the approximation outlined in section 6.2, we need to know c_1 and c_2 simultaneously, since the approximation for S_2 involves c_1+c_2 and that for S_1 depends on the mean queue length of S_2. Although an iterative method may be used to solve for both quantities, we settle for a simpler, albeit slightly less accurate, approach by assuming that S_2 has infinite waiting capacity. This, in fact, is the original assumption used by Koizumi *et al.*, which we modified in section 6.2 by enforcing S_2 to be a finite capacity $M/M/c_2/(c_1+c_2)$ queue, as this modification improved the accuracy. Moreover, since we only have analytical results for the wait time distribution of an $M/M/(\cdot)$ queue, we will use the original method of Koizumi *et al.*, and in doing so let S_2 be an $M/M/c_2$ queue.

Once we decompose the tandem system into two separate queues, we can use equation (5.23) to obtain the pdf $h_a^{(i)}(t)$ of the wait time for any unit acquiring service. Then, from the cumulative distribution function

$$H_i = \int_0^{T_i} h_a^{(i)}(t)\, dt, \tag{7.1}$$

the percentage of units acquiring service in S_i within a given time limit T_i can be determined. For each station S_i, this quantity depends on the number of servers c_i and the reneging parameter α_i, both of which are unknown a priori. However, it must be that for any chosen c_i and α_i, the percentage of reneged units is equal to κ_i, as given by the database. In other words, for each station there are two unknowns c_i and α_i that must be determined from two constraints H_i and κ_i. Mathematically, this can be written as a system of two nonlinear equations, which upon inversion yield the unknown parameters:

$$\begin{aligned} H_i &= H_i(c_i, \alpha_i), && (7.2) \\ \kappa_i &= \kappa_i(c_i, \alpha_i)\,. && (7.3) \end{aligned}$$

For easier reference, we note that

$$\begin{aligned} H_1 &= 90\%, & H_2 &= 80\%, \\ T_1 &= 1 \text{ hour}, & T_2 &= 6 \text{ hours}, \\ \kappa_1 &= 2.39\%, & \kappa_2 &= 2.32\%. \end{aligned}$$

The system (7.2)–(7.3) can be solved for c_i and α_i by standard root-finding methods. The approach we take is as follows: We start with an initial guess of c_i, for $i = 1, 2$, so that an approximate value of the queue reneging parameter α_i can be obtained via equation (5.25) using the estimated reneging probability κ_i. This allows us to obtain $H_i(c_i, \alpha_i)$. If this value is lower than the target access rate H_i, then c_i is increased[1], and vice versa, always rounding c_i to the nearest integer. The process is repeated until both equations (7.2) and (7.3) are best satisfied (clearly, since c_i is an integer, we can only hope to get results that are close to H_i). We now proceed to determine the parameters of the decomposed queueing system.

[1] Since access rate increases with the number of available beds, H_i is an increasing function of c_i.

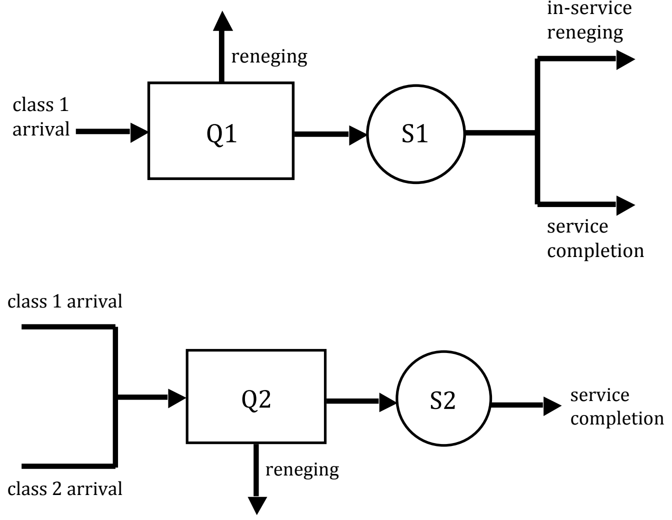

Figure 7.3: Decomposed S_1 and S_2 queues

From the parameter values specified in the previous section, we estimate the external arrival rates to each station as

$$\begin{aligned}\lambda_1 &= \frac{N_1}{365 \text{ days}} = 2.17 \text{ days}^{-1},\\ \lambda_2 &= \frac{N_2}{365 \text{ days}} = 27.34 \text{ days}^{-1}.\end{aligned}$$

Now recall that the queue leading to S_2 is composed of the new arrivals and the blocked units in S_1. The former process is Poisson with parameter λ_2, while the latter, due to the output process theorem (see section 2.8), is Poisson with parameter

$$\lambda_1' = \lambda_1(1-\kappa_1') = 1.83 \text{ days}^{-1},$$

since S_1 is treated as an $M/M/c_1$ queue, and only a fraction $1-\kappa_1' = 84.2\%$ of the units arrived to S_1 finish their service without reneging during service. Thus, by the additive property of the Poisson process, the effective arrival rate to S_2 is

$$\lambda_2' = \lambda_1' + \lambda_2 = 29.23 \text{ days}^{-1}.$$

In addition, since S_2 does not experience blocking or reneging, the service rate can directly

c_2	α_2	6 hr access	mean queue length	mean wait time
	10^{-3} hrs^{-1}	%		hrs
202	1.20	32.6	23.1	19.0
203	1.82	44.0	15.5	12.7
204	2.62	54.2	10.8	8.84
205	3.69	63.1	7.66	6.29
206	5.13	71.0	5.51	4.53
207	7.11	79.5	3.97	3.26
208	9.91	83.6	2.85	2.34
209	14.0	88.5	2.02	1.66
210	20.2	92.6	1.40	1.15
211	30.0	95.7	.943	.774
212	46.7	98.0	.606	.497

Table 7.1: Approximate performance measures for S_2

be computed from the average length of stay as

$$\mu_2 = \frac{1}{W_2'} = 0.143 \text{ days}^{-1}.$$

Knowing these parameters, we can now use the procedure discussed earlier to determine c_2 and α_2 from H_2 and κ_2. Table 7.1 shows the results. More specifically, it illustrates how the percentage of units receiving a server within 6 hours varies with the number of servers. It can be seen that the server requirement to achieve an 80% access to S_2 is

$$c_2 = 207$$

and the corresponding reneging rate is

$$\alpha_2 = 0.171 \text{ days}^{-1}.$$

We now turn our attention to S_1. Firstly, from equation (5.29) we have that

$$\kappa_1' = \alpha_1' W_1', \tag{7.4}$$

from which we obtain the in-service reneging rate as

$$\alpha_1' = 0.0356 \text{ days}^{-1}.$$

c_1	α_1	1 hr access	mean queue length	mean wait time
	10^{-3} hrs^{-1}	%	10^{-2}	hrs
10	.248	23.7	844	93.3
11	.967	53.8	222	24.6
12	2.74	73.8	78.4	8.68
13	7.68	86.7	28.0	3.09
14	25.7	94.5	8.36	.925
15	173	99.3	1.24	.137

Table 7.2: Approximate performance measures for S_1

If there were no blocking in S_1, we could write a similar relationship for the service rate, or the reciprocal of the mean length of stay:

$$1 - \kappa_1' = \mu_1 W_1'. \tag{7.5}$$

However, due to blocking, the mean length of stay is given by the mean service time μ_1^{-1} plus the mean duration of blocking, which can be approximated by $(\lambda_1'/\lambda_2')W_2$. This is because W_2 is the mean waiting time to enter S_2, but since blocked class 1 units have priority over class 2 units, approximately only a fraction λ_1'/λ_2' of this wait time can be attributed to blocking – this fraction is an estimate of the ratio of blocked class 1 units to all units entering S_2. Therefore, to incorporate the blocking effect, equation (7.5) needs to be modified to

$$1 - \kappa_1' = \frac{W_1'}{\mu_1^{-1} + W_2(\lambda_1'/\lambda_2')} . \tag{7.6}$$

From table 7.1, we find that when $c_2 = 207$

$$W_2 = 0.136 \text{ days},$$

so that solving equation (7.6) for μ_1 yields

$$\mu_1 = 0.190 \text{ days}^{-1} .$$

This is the actual service rate in S_1, which is to be used in the simulation. However, in using Method 4 of section 6.2, we only need the effective service rate, which has blocking time incorporated in it. As discussed earlier, this is given by

$$\mu_{\text{eff}} = W_1'^{-1} = 0.225 \text{ days}^{-1} .$$

Using the same approach as before, we obtain c_1 and α_1 so that equations (7.2) and (7.3) are best satisfied. From the results in table 7.2 we can see that the required number of servers in S_1 is

$$c_1 = 14$$

and the corresponding reneging rate is

$$\alpha_1 = 0.616 \text{ days}^{-1} .$$

An interesting observation is that the in-queue reneging rate α_1 is about 17 times larger that the in-service reneging rate α'_1. In queueing theory terminology, this means that customers who are being served have a much higher patience than those waiting in line. The implication of this to the hospital setting is that critically ill patients who are already in the ICU under treatment have a much higher survival tolerance compared to those who are in ED (the patience of customers at a service station is analogous to the survival tolerance of patients at a treatment facility). This phenomenon is known as the *Golden Hour* in medicine, which refers to the period of a few minutes to several hours following a trauma during which the possibility of death is greatest, and consequently the chances of survival are greatest if the victims receive care within a short period of time after the accident [38]. Our results regarding the differences between in-queue and in-service reneging rates are in agreement with this observed phenomenon.

Having estimated the reneging parameters α_1, α_2, and α'_1, in addition to the service rate μ_1, we next use simulation to refine our estimated server requirement of $c_1 = 14$ and $c_2 = 207$.

7.3 Simulation

To simulate the queueing system described in the previous section, we use the SimEvents package from MATLAB, which is an extension of the Simulink software for doing Discrete-Event Simulation (DES). In general, DES is a method of modelling in which state variables describing the system under study change at discrete points in time. Two main approaches used in advancing the simulation clock are *next-event time advance* and *fixed-increment time advance*. It is the first approach which is more commonly used in simulation software, including SimEvents.

In next-event time advance, the simulation clock is initialized to zero and the times of occurrence of future events are determined (examples of future events in our model are arrivals, service completions, and reneging of units in queue or in service). The simulation clock is then advanced to the first of these future events (in chronological order), at which point the state of the system along with the list of future event times is updated to account for the fact that an event has occurred. The simulation clock is then advanced to the time of the next event, and the process continues until eventually some prespecified stopping condition is satisfied. Since all state changes occur only at event times, periods in which no event occurs are skipped over by jumping the clock from one event time to the next. On the contrary, fixed-increment time advance does not skip over these inactive periods, resulting in lengthened computational time [22].

Since from the analytical calculations we have an estimate of the number of required servers in S_1 and S_2, namely $c_1 = 14$ and $c_2 = 207$, we performed the simulations for a range of parameters close to these initial guesses. Two-year simulations were used, with the first year ignored as the transient period. Performance measures were then computed for each patient who arrived during the last year. For each stream of arrivals, we obtained the percentage of units who acquire service within the required time limit (one hour for class 1 units, and 6 hours for class 2 units). Then, for each choice of c_1 and c_2, we ran the simulation 50 times and averaged the results. This method of running relatively short simulations many times is advocated by Pawlikowski [26], who states that in order to infer the population mean from a sample of data obtained from simulation, single long simulations must be replaced by many short ones; this is referred to as the method of independent replications and is meant to avoid the problem of correlation between successive data points, which is magnified during long simulation runs.

Table 7.3 shows the simulation results for various values of c_1 and c_2 close to the estimated counts. From the table we can see that $c_1 = 14$ and $c_2 = 208$ provides access of 90.7% to the S_1 in 1 hour and 82.3% to the S_2 in 6 hours. These results are in close agreement with those obtained from the analytical approximation. The reason for this is that there

c_1	c_2	% to S_1 in 1 hr	mean wait hrs for S_1	% to S_2 in 6 hrs	mean wait hrs for S_2
13	206	83.8	2.54	70.8	4.43
13	207	83.5	2.57	78.2	3.03
13	208	83.0	2.88	83.1	2.21
13	209	81.6	2.88	86.8	1.63
13	210	82.8	2.67	90.9	1.07
14	206	91.6	.700	69.6	4.64
14	207	90.8	.766	76.6	3.29
14	208	90.7	.800	82.3	2.31
14	209	90.3	.815	86.5	1.66
14	210	91.6	.721	91.1	1.06
15	206	96.3	.063	70.5	4.41
15	207	96.5	.062	80.3	2.73
15	208	96.9	.046	83.3	2.19
15	209	96.3	.058	87.5	1.53
15	210	96.5	.057	90.7	1.10

Table 7.3: Performance measures for S_1 and S_2 obtained via simulation

are sufficient beds in the MU to reduce the blocking probability to almost zero, hence the interdependence of the two stations can effectively be ignored in this case. This fact is also evident from table 7.3, where we observe that the variation in c_2 appears to have almost no effect on the percentage of access to S_1 in 1 hour (the small variation is presumably due to the random nature of the simulation). Similarly, the percentage of access to S_2 in 6 hours appears unaffected by the variation in c_1. These results were expected from the analytical calculations, where we estimated the average blocking time for class 1 units to be $(\lambda_1'/\lambda_2')W_2 = 12.2$ minutes, which compared to the average length of stay in S_1, $W_1' = 4.44$ days, is negligible. The main reason that a significant blocking effect is not observed in our model is that we attributed a higher priority to class 1 units; if both units had had the same priority, then the blocking time would have been $W_2 = 3.26$ hours, much larger than 12.2 minutes.

To obtain confidence intervals for the simulation estimates, define the random variable

$$X_i = \% \text{ of units receiving service in } S_1 \text{ within 1 hour in the } i\text{th experiment.}$$

Now, since each run of the experiments is performed using different seeds for the random number generators, the X_i can be viewed as independent and identically distributed random variables. From the central limit theorem in statistics, it follows that as n increases, the sample average of these random variables $\overline{X}(n) = (X_1 + \cdots X_n)/n$ approaches the normal distribution with mean $\eta = E\{X_i\}$ and standard deviation $\sigma/\sqrt{n}$. Now, since the sample variance s^2 converges to the true variance σ^2 as n gets large, the theorem implies that for large values of n, the sample mean $\overline{X}(n)$ is approximately distributed as a normal random variable with mean η and standard deviation $s/\sqrt{n}$. The quantity $s/\sqrt{n}$ is referred to as the standard error of the mean, which is defined as the standard deviation of the sample mean estimate of a population mean.

Since in our work $n = 50$, and values of $n > 30$ are usually considered large enough for the central limit theorem to hold, we can use the normal distribution to obtain confidence intervals for our estimates. In particular, for $c_1 = 14$ and $c_2 = 208$, the simulation produced a 90.7% access to S_1 in 1 hour with standard error of 0.5%, which gives $(89.7\%, 91.7\%)$ as the 95% confidence interval. Similarly, we find $(80.9\%, 83.7\%)$ as the 95% confidence interval for the percentage of units receiving service in S_2 within 6 hours. Note that since the original values of our parameters obtained from the database have measurement errors in them, the confidence intervals are estimates also.

7.4 Conclusion

In this project we developed a queueing network model with blocking and reneging to study how the wait times in the ED are influenced by the number of available beds in the ICU and the MU. The in-queue reneging phenomenon is due to patients' death or treatment completion in the ED, while the in-service reneging refers to deaths that occur in the ICU when a patient is already under treatment. Since the ICU and the MU have multiple beds, we studied the $M/M/c$ queue with reneging and obtained a relationship between the percentage of reneged patients and the reneging parameter in addition to finding the wait time distribution for the units receiving service. However, as these results could not readily be applied to the tandem ICU-MU queueing system, we used approximate methods to decompose the system into two independent multi-server queues. This approximation was then used to obtain estimates for the required number of beds in the ICU and the MU so that specified targets of access were met. Subsequently, guided by the estimated bed counts, simulation produced more refined estimates. More concretely, by requiring a TTA of 90% within 1 hour for the ICU and 80% within 6 hours for the MU, we found that the required number of beds is approximately 14 in the ICU and 208 in the MU for the test hospital under consideration.

These results were based on the parameters obtained from the 2006-07 database. Thus, it is by no means a prediction of what the future number of beds should be. This project is part of a larger one aimed at estimating bed requirements in B.C. acute care hospitals, given forecasted population demographics. To achieve that goal, several modifications and additions need to be introduced into this model, which we discuss next.

7.5 Future Work

As mentioned in the Introduction, in this project we have only considered the emergency stream of patients. There are also elective admissions and transfers from other hospitals that need to be taken into account in order to accurately determine the required number of beds. In addition to including these two streams of patients, the variations in the arrival rate throughout the day or even the week need to be considered. In this project, we assumed a constant arrival rate throughout the day, which is not a reasonable assumption, for there are large variations in the 24-hour period. In the next phase of the project, we will use a piecewise constant function for the arrival rate which is averaged over six four-hour periods in a day. Moreover, we need to consider how accurate it is to assume that inter-arrival and service times are exponentially distributed, and possibly consider more general distributions that better fit the empirical data. Finally, we further need to investigate the policy by which doctors decide on the priority of patients entering the MU from the ICU and the ED. As we saw, giving priority to ICU patients in effect eliminates blocking. Suggesting this policy to the hospital management could be a possibility. On the other hand, if this assumption is unrealistic, a better decision process needs to be implemented to reflect more accurately the transfer of patients from the ICU to the MU.

Appendix A

$M/M/1$ Queue with Reneging

Ancker and Gafarian [3] analyzed the $M/M/1$ queue with reneging, but with the additional complexity that an arriving unit, upon finding n units already in the system, balks (does not join the queue) with probability n/N, where N is the maximum number of units allowed in the system. It can be seen that this queue is equivalent to the $M/M/1$ queue with reneging only in the limit as $N \to \infty$, and so their result can be used as a check for equation (5.22). They showed that the wait time distribution of those units who join the queue and successfully acquire service is given by

$$g_a^{(N)}(t) = \lambda z^{\gamma}(1-z)^{N-1} e^{-(\mu+\alpha)t} \frac{\left[1+\eta(1-e^{-\alpha t})\right]^{N-2}}{N B_z(\gamma+1, N-1)}$$

where

$$\begin{aligned} \eta &= \lambda/N\alpha, \\ \gamma &= \mu/\alpha, \\ z &= \eta/(1+\eta), \end{aligned}$$

and $B_z(\gamma, N)$ is the incomplete Beta function given by

$$B_z(\gamma, N) = \int_0^z \gamma^{\gamma-1}(1-\gamma)^{N-1}\, d\gamma\,.$$

For our purposes, we would like the behaviour without any balking constraint; in other words, we are looking for $g_a^{(N)}(t)$ in the limit $N \to \infty$, so that there is no limit on the maximum number of customers in the system.

As in chapter 5 let us define $\delta = \lambda/\alpha$, so that $z = \delta/(N+\delta)$. Note that $\lim_{N\to\infty} z = 0$. The expression for $g_a^{(N)}(t)$ involves the term $(1-z)^{N-1}$ which in the limit of $N \to \infty$ gives

$$\begin{aligned}
\lim_{N\to\infty} (1-z)^{N-1} &= \lim_{N\to\infty} (1-z)^{-1}\left(1 - \frac{\delta}{N+\delta}\right)^N \\
&= \frac{\delta}{N+\delta}\left(1 - \frac{\delta}{N}\right)^N \\
&= e^{-\delta}\,.
\end{aligned}$$

Similarly, since $\lim_{N\to\infty} \eta = 0$ we get that

$$\begin{aligned}
\lim_{N\to\infty} \left[1 + \eta\left(1 - e^{-\alpha t}\right)\right]^{N-2} &= \lim_{N\to\infty} \left[1 + \eta\left(1 - e^{-\alpha t}\right)\right]^{-2}\left[1 + \frac{\delta\left(1 - e^{-\alpha t}\right)}{N}\right]^N \\
&= \exp\left\{\delta\left(1 - e^{-\alpha t}\right)\right\}\,.
\end{aligned}$$

Now, we compute the limiting behaviour of the incomplete Beta function. By changing the variable of integration to $x = t(N-2)$ and noting that $\lim_{N\to\infty} zN = \delta$ we find that for large values of N

$$\begin{aligned}
B_z(1+\gamma, N-1) &= \int_0^z t^\gamma (1-t)^{N-2}\, dt \\
&= \int_0^{z(N-2)} \left(\frac{x}{N-2}\right)^\gamma \left(1 - \frac{x}{N-2}\right)^{N-2} \frac{1}{N-2}\, dx \\
&= \frac{1}{(N-2)^{1+\gamma}} \int_0^{z(N-2)} x^\gamma \left(1 - \frac{x}{N-2}\right)^{N-2} dx \\
&\sim N^{-(1+\gamma)} \int_0^\delta x^\gamma e^{-x}\, dx \\
&= N^{-(1+\gamma)} \Gamma_l(\gamma+1, \delta)\,.
\end{aligned}$$

This result allows us to find the following limiting behaviour as $N \to \infty$:

$$\begin{aligned}
\lim_{N\to\infty} \frac{z^{\gamma}}{NB_z(\gamma+1, N-1)} &= \lim_{N\to\infty} \frac{\left(\frac{\delta}{\delta+N}\right)^{\gamma}}{NB_z(\gamma+1, N-1)} \\
&= \lim_{N\to\infty} \frac{\delta^{\gamma}}{N^{1+\gamma}B_z(\gamma+1, N-1)} \\
&= \lim_{N\to\infty} \frac{\delta^{\gamma}}{\Gamma_l(\gamma+1, \delta)}.
\end{aligned}$$

Putting these results together, we obtain the limiting behaviour of $g_a^{(N)}(t)$ as $N \to \infty$:

$$\lim_{N\to\infty} g_a^{(N)}(t) = \frac{\alpha\delta^{\gamma+1}}{\Gamma_l(\gamma+1, \delta)} \exp\{-(1+\gamma)\alpha t - \delta e^{-\alpha t}\}.$$

This is the same result as equation (5.22), which was obtained from the $M/M/c$ queue with reneging in the special case of $c = 1$.

Bibliography

[1] J. Abate, *Numerical inversion of Laplace transforms of probability distributions*, ORSA Journal on Computing, 1995, Vol. 7, No. 1, pp. 36–43.

[2] I. Adan and J. Resing, *Queueing Theory*, Online lecture notes, Department of Mathematics and Computing Science, Eindhoven University of Technology, 2001. `http://www.win.tue.nl/∼iadan/queueing.pdf`

[3] C. J. Ancker and A. V. Gafarian, *Some queueing problems with balking and reneging. I*, Operations Research, 1963, Vol. 11, No. 1, pp. 88–100.

[4] S. Balsamo, V. De Nitto Personé, R. Onvural, *Analysis of queueing networks with blocking*, Kluwer Academic Publishers, Massachusetts, 2001.

[5] P. Bastani, B. Ramadanovic, A. R. Rutherford, L. Vertesi, Y. Wang, *Modelling acute care in hospitals, phase 2 — Model description*, Report for the British Columbia Ministry of Health, Complex Systems Modelling Group, 2009.

[6] S. K. Bose, *An Introduction to Queueing Systems*, Kluwer Academic/Plenum Publishers, New York, 2002.

[7] O. J. Boxma, A. G. Konheim, *Approximate analysis of exponential queueing systems with blocking*, Acta Informatica, 1981, Vol. 15, No. 1, pp. 19–66.

[8] A. M. de Bruin, A. C. van Rossum, M. C. Visser, G. M. Koole, *Modelling the emergency cardiac in-patient flow: an application of queueing theory*, Health Care Management Science, 2006, Vol. 10, No. 2, pp. 125–137.

[9] J. K. Cochran, A. Bharti, *Stochastic bed balancing of an obstetrics hospital*, Health Care Management Science, 2006, Vol. 9, No. 1, pp. 31–45.

[10] R. B. Cooper, *Introduction to Queueing Theory*, The Macmillan Company, New York, 1972.

[11] N. M. van Dijk, *Queueing Networks and Product Forms: a Systems Approach*, John Wiley & Sons Ltd., Chichester, 1993.

[12] E. el-Darzi, C. Vasilakis, T. Chaussalet, P.H. Millard, *A simulation modelling approach to evaluating length of stay, occupation, emptiness, and bed-blocking in a hospital geriatric department*, Health Care Management Science, 1998, Vol. 1, No. 2, pp. 143–149.

[13] P. Godfrey, *Rational polynomial curve fitting*, MATLAB Central, 2006, File ID: # 11197, `http://www.mathworks.com/matlabcentral/fileexchange/11197`.

[14] F. Gorunescu, S. I. McClean, and P. H. Millard, *A queueing model for bed-occupancy management and planning of hospitals*, The Journal of the Operational Research Society, 2002, Vol. 53, No. 1, pp. 19–24.

[15] L.V. Green, *How many hospital beds?*, Inquiry, 2002, Vol. 39, No. 4, pp. 400–412.

[16] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Third Edition, Wiley-Interscience, 1998.

[17] X. Huang, *A planning model for requirement of emergency beds*, Journal of Mathematics Applied in Medicine and Biology, 1995, Vol. 12, No. 3-4, pp. 345–353.

[18] J. R. Jackson, *Networks of waiting lines*, Operations Research, 1957, Vol. 5, No. 4, pp. 518–521.

[19] S. Jain and M. Smith, *Open finite queueing networks with $M/M/C/K$ parallel servers*, Computers and Operations Research, 1993, Vol. 21, No. 3, pp. 297–317.

[20] N. Koizumi, E. Kuno, T. E. Smith, *Modelling patient flows using a queueing network with blocking*, Health Care Management Science, 2005, Vol. 8, No. 1, pp. 49–60.

[21] R. Korporaal, A. Ridder, P. Kloprogge, and R. Dekker, *An analytical model for capacity planning of prisons in the Netherlands*, Journal of the Operational Research Society, 2000, Vol. 51, No. 11, pp. 1228–1237.

[22] A. M. Law, W. D. Kelton, *Simulation Modelling and Analysis*, McGraw-Hill Companies Inc, USA, 2000, Third Edition.

[23] J.D.C. Little, *A proof of the queueing formula $L = \lambda W$*, Operations Research, 1961, Vol. 9, No. 3, pp. 383–387.

[24] S. McClean and P. Millard, *Modelling in patient bed usage in a department of geriatric medicine*, Methods of Information in Medicine, 1993, Vol. 32, No. 1, pp. 79–81.

[25] P.M. Morse, *Queues, Inventories and Maintenance*, Wiley, New York, 1958.

[26] K. Pawlikowski, *Steady-state simulation of queueing processes: survey of problems and solutions*, ACM Computing Surveys, 1990, Vol. 22, No. 2, pp. 123–170.

[27] H. G. Perros, *Queueing Networks with Blocking*, Oxford University Press: Oxford, New York, 1994.

[28] M. C. Pike, D. M. Proctor, and J. M. Wyllie, *Analysis of admissions to a casualty ward*, British Journal of Preventive and Social Medicine, 1963, Vol. 17, No. 4, pp. 172–176.

[29] N. U. Prabhu, *Foundations of Queueing Theory*, International Series in Operations Research & Management Science, Kluwer Academic Publishers, Massachusetts, 1997.

[30] Bojan Ramadanovic, CSMG, IRMACS, *unpublished notes.*

[31] A. R. Rutherford, L. Vertesi, R. Ferguson, W. Hare, J. Li, K. Vasarhelyi, and Y. Wang, *Modelling acute capacity in hospitals, phase 1 — Modelling methodology*, Report for the British Columbia Ministry of Health, Complex Systems Modelling Group, 2008.

[32] K. Siddharthan, W. J. Jones, J. A. Johnson, *A priority queueing model to reduce waiting times in emergency care*, International Journal of Health Care Quality Assurance, 1996. Vol. 9, No. 5, pp. 10–16.

[33] W. J. Stewart, *Introduction to the Numerical Solution of Markov chains*, Princeton University Press, Princeton, 1994.

[34] Les Vertesi, CSMG, IRMACS, *private communication.*

[35] E. N. Weiss, J. O. McClain, *Administrative days in acute care facilities: a queueing-analytic approach*, Operations Research, 1986, Vol. 35, No. 1, pp. 35–44.

[36] D. V. Widder, *Advanced Calculus*, Dover Publications, 1961, Second Edition.

[37] D. J. Worthington, *Queueing models for hospital waiting lists*, The Journal of the Operational Research Society, 1987, Vol. 38, No. 5, pp. 413–422.

[38] American College of Surgeons, *ATLS Advanced Trauma Life Support Program for Doctors*, American College of Surgeons, 2008, Eighth Edition.