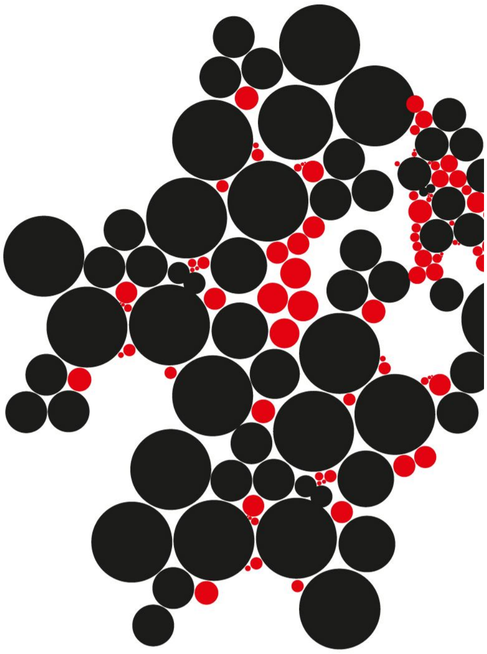
MASTER THESIS



LOSS NETWORKS WITH RESERVATION

N.M. van de Vrugt

APPLIED MATHEMATICS
STOCHASTIC OPERATIONS RESEARCH

EXAMINATION COMMITTEE
Prof. dr. R.J. Boucherie
Dr. N. Litvak
Dr. G. Still

UNIVERSITY OF TWENTE.

AUGUST 19, 2011

Loss Networks with Reservation

Final report

N.M. van de Vrugt

Supervisors: Prof.dr. R.J. Boucherie
Dr. N. Litvak

Faculty Electrical Engineering, Mathematics and Computer Science
Department Applied Mathematics
Chair Stochastic Operations Research
Drienerlolaan 5
Postbus 217
7500 AE Enschede
e-mail: info@ewi.utwente.nl

August 2011

Summary

Like many other health care institutes, rehabilitation centre “Het Roessingh” (Enschede, The Netherlands) has difficulties to satisfy the demand for rehabilitation care. In the scheduling department of this rehabilitation centre holistic (entire centre instead of single departments) and long-term view is lacking. This can easily result in suboptimal planning and may lead to suboptimal quality of care due to unsynchronized appointments at different departments. Het Roessingh would like to utilize clinical pathways for both enlarging the scheduling horizon and enforcing the treatments for one patient at different departments to be in the same time period. The latter implies that a patient can be forced to wait several weeks if this results in all treatments starting in the same week. Since rehabilitation care is typically non-emergency care, letting patients wait (less than a specified norm) is allowed in Het Roessingh.

The fundamental mathematical research in this report originated from the rehabilitation care problem described above. We introduce a so-called *reservation model* and examine its properties. In this queueing model patients arrive to a tandem queue of an infinite server queue (the reservation queue) and a loss queue. The tandem network has an exceptional blocking rule; when a patient arrives to the reservation queue, the service requirements at both queues are drawn from the appropriate distributions and it is checked whether there are sufficient resources for the new patient at the second queue in the time interval the new patient requires service at the second queue. If resources are insufficient, the patient is blocked and lost. This network represents a loss network in which resources can be claimed a random time in advance.

In this research we first derive the stationary distribution of the reservation model. Finding the stationary distribution appeared to be difficult when we tried solving the Kolmogorov differential equations. Therefore we derived the stationary distribution for a deterministic single server reservation model by means of a renewal theory argument. With this result we prove the reservation model does not have a product-form stationary distribution.

Besides these analytic results, we performed an extensive simulation study. In this study we compared the probability an arriving patient is blocked (hereafter: blocking probability) in the reservation model with the blocking probability in the ordinary loss queue. We found that for deterministic reservation and/or service time the reservation model has a blocking probability greater than or equal to the blocking probability of the loss queue without reservation. We proved this claim for all capacity constraints of the second queue. For exponential reservation and service requirements the results were the other way around; reservation results in less patients being blocked. These results were found for several capacity constraints of the loss queue, but we only give an outline of the proof for capacity 1.

Preface

When I started at the University of Twente five years ago, I could only hope I was writing this Preface at this moment. I was very nervous I would appear terrible at Mathematics and thought I would soon discover I lacked the quality to finish my courses. At that moment I had never thought I would accept the great opportunity of joining the Honours Programme during my second year. In this programme I formulated my research proposal for this final project and (hopefully) for my PhD-project. It took a while for me to realize that graduating on your own research proposal is quite unique in the field of Mathematics and I am very gratefull for all support I got during the Honours Programme.

In a Preface of a final project, one usually looks back at his/her student days. Looking back, I can say I had a really good time meeting new friends and living “on my own”. Several weeks before my first lecture I found a room in the centre of Enschede, where I lived for almost five years together with several girls who became my close friends. They almost refused me to apply for the room because I was about to become a math student, but fortunately they did invite me for meeting them and eventually I got the room.

When I signed in for my studies Applied Mathematics, I can not deny I was affraid my entire class would consist of nerds. I was happy to find out my class was really nice and I am sure I will keep contact with most of them (especially the girls). Whether they (or should I say *we*) are nerds? Let’s leave that question for someone else...

During my Bachelor studies I appriciated the fact we got introduced to many different kinds of mathematics. I can not say I liked every subject, but most of them were very interesting. For my final project (for the Bachelor title) I worked in a group of six at finding the optimal pension investment strategy. Obviously, this was an assignment of the chair Financial Engineering. I had chosen this assignment because I wanted to do one different subject before I would spend the two years of my Masters at queueing theory. It was very enjoyable with that many groupmates and we had a great time. On the other hand it appeared hard to work together with that many people, since we spend more time on discussing which method we should use than on actually solving our problem. I can say I learned a lot from this project.

As said before, I had made up my mind about choosing the Stochastic Operations Research Masters Track fairly early. In the second year of my Masters I did an internship at Witteveen + Bos in Deventer. The project I worked on was already promising before I started my internship, so for me it was hardly a surprise the results were really good. I have enjoyed working with my colleagues there and still enjoy all the publicity that resulted from my internship.

Since I had achieved that much in the three months of my internship, the first few months of my final project were a bit disappointing. Perhaps because I was envolved during the first stage of the research proposal, I had hoped for more results. The model Richard and I had invented, appeared not only mathematically non-trivial but it also was very complicated to derive analytical results. I can not say I do not like difficult problems, but (as usual at final projects) there were moments I wished I had never written the research proposal at all. Fortunately after five months the results described in this report began to take shape. Just in time I started to like my research again, since I am hired at the University of Twente to do a PhD-project on my final project!

The remainder of this Preface I would like to spend on thanking my supervisors, Nelly and Richard, for their help, patience and good ideas during my final project. Also I would like to thank my officemate Niek and my friends and family for their love and support. Finally, I wish the reader joy with reading my report!

Maartje van de Vrugt
maartjevdvrugt@gmail.com
Enschede, August 19, 2011.

Contents

Chapter 1

Introduction

Nowadays people live longer then they did a few hundred years ago, although the expected number of years people live without a chronic disease is decreasing [9]. Moreover the Netherlands cope with relatively many elderly inhabitants and the population of our country will be growing till at least 2050 [28]. As a consequence the pressure on our health care system will grow substantially.

Like many other branches, rehabilitation centres are struggling to keep up with the growing demand. In many cases there is no money available to increase resources. As a consequence institutes have to provide care more efficiently.

This research is one in a large set of mathematical researches that originated in health care logistics and scheduling. For literature studies on this topic, see for example [21] and [36]. Few of these researches focus on rehabilitation care while the kind of scheduling problems differ from those in hospitals. Many patients that access a rehabilitation centre, have to receive treatment from more then one department in the institute and visit departments more than once. Most of the patients follow treatment schemes that can be predicted very accurately.

Despite the fact that treatment schemes can be predicted accurately, the scheduling department of Het Roessingh (a rehabilitation centre in Enschede, The Netherlands) only makes schedules for one week. The department (and staff and probably the patients) would like to increase this scheduling horizon.

Another problem at this rehabilitation centre is that it is fully occupied nearly all the time and it is difficult to schedule all patients within the Treek-norm. The Treek-norm is a legally determined upper bound on the access time of a patient [10]. The planning department is forced to cancell appointments of other patients in order to fit in the arriving patients when they cannot access Het Roessingh within the norm. Long-term scheduling is therefore impossible and the quality of care could decrease when treatments are cancelled.

When a patient at Het Roessingh has to receive treatment at multiple departments, the requests for appointments are treated separately at the planning department. Since all departments have different occupation rates, it could occur that a patient has received all treatment at one department and has to start treatment at another. This way the therapists cannot gear the treatments to eachother, which may lead to suboptimal quality of care.

Concluding it can be said that the lack of holistic and long-term view at Het Roessingh can easily result in suboptimal planning and may lead to suboptimal quality of care. The rehabilitation centre wants to tackle this problem by utilizing so-called clinical pathways (CPs). CPs are developed from scheduling techniques used in the industries, for example the Critical Path Method and the Program Evaluation Review Technique [33]. Because nowadays the CPs are used in many institutes individually, there are many different definitions attached to this term (e.g. see [13]). Het Roessingh has defined clinical pathways for each type of disorder as a set of different treatments. For some rare disorders the CPs could be created specially for one or two patients, while others are frequently used. The frequently used pathways contain additional information on the required number of treatments at each department. The therapists always try their best to finish the treatment within the specified number of appointments. When a patient enters Het Roessingh he/she is labelled with a pathway, which immediately defines the treatment plan. Occasionally, when the treatment plan does not have the intended result a patient has to be relabelled.

As said before, Het Roessingh would like to utilize the clinical pathways for both enlarging the scheduling horizon and enforcing the treatments of one patient at different departments to be in the same time period. The latter implies that a patient can be forced to wait several weeks if this results in all treatments starting in the same week. Since rehabilitation care is typically non-emergency care, letting patients wait (less than the specified norm) is allowed in Het Roessingh. Before testing scheduling with CPs in real-life, the rehabilitation centre would like to know the impact of this way of scheduling on for example the average access time and utilization rate of the system. This question was the inspiration of the research question that is given in the following section. In the final section of this chapter a brief overview of the structure of this report is given.

1.1 Research question

There are many interesting questions that could be topic of this final project since scheduling with CPs has impact on the access times, blocking probability, utilization rate of the system, etc. Moreover we could look at the actual scheduling or take a more holistic viewpoint. We choose to focus on the effect of reservation of resources on the blocking probability. For Het Roessingh the blocking probability is an important performance measure and mathematically reservation is an interesting feature of the system, since then the first-come-first-serve principle does not need to be valid. When we can model the loss network with reservation, we could easily describe patients that require multiple appointments in a time interval and this makes the possibility of reservation the most interesting topic of this final project.
After a literature review (as described in the next chapter) we concluded that the best way to model CPs with reservation is with a loss network. Therefore we define the following research question:

What is the effect of reservation of resources on the blocking probability in a loss network?

We will find an answer to this question in two different ways. First we will formally describe the queueing model and try to find the stationary distribution. With this distribution we can express the blocking probability analytically. For the ordinary loss network the analytic blocking probability is already known and with this probability we can determine the effect of reservation analytically.
The second way to answer the research question is by performing an extensive simulation study. We will build a discrete event simulation representing an ordinary loss network and one with reservation and compare the blocking probabilities of both models.

1.2 Overview of this report

In this report we first present a literature review on the effect of reservation on the blocking probability. Hereafter the mathematical model that is central in this research will be described formally. Several chapters following the model, contain the analytical and numerical results. The final chapters are the conclusion and discussion.

Chapter 2

Literature review

A recent article by Schemmelpfeng, Helber and Kasper ([32]) addresses the importance of improving scheduling techniques in rehabilitation centres. In this article the authors give a good literature review and notice that "compared to the substantial number of publications dealing with planning and scheduling in acute hospitals, only a few papers address the scheduling problems in rehabilitation hospitals". In order to create a scheduling tool that can provide schedules for the entire rehabilitation centre, in [32] formal mixed-integer linear programs are developed.

Research on clinical pathways so far has concentrated mainly on the optimization of the schedules and the use of capacity on a single department. Most literature on clinical pathways generally describes case-studies and debates whether CPs optimize planning on single departments. The remainder of the literature is about how to develop, implement and measure clinical pathways and occasionally on the interference between two departments [36].

It appears clinical pathways are not used for a holistic view (for example the viewpoint of an entire hospital) of health care institutes. This is striking because from the industry it is generally known an atomistic view (for example the viewpoint of each department separately) is inefficient. Since CPs originated in the industries from the Program Evaluation Review Technique (PERT) and the Critical-Path Method (CPM), the applicability of these existing optimization methods on scheduling clinical pathways and the possibility of evaluating the impact of CPs scheduling with these and other scheduling methods, will be evaluated first. Hereafter the possibilities with queuing theory are investigated.

2.1 Scheduling methods

In the industries PERT is frequently used for optimizing networks of events and activities. In those networks it is given that certain events have to occur before other events, i.e. certain activities have to be finished before others can be started. As a consequence a critical path inside the network can be obtained by sequencing the most time-consuming activities that have to be done successively. This critical path has to be scheduled first and all other activities can be scheduled on the basis of this path. By means of the earliest and latest expected starting time of the activities, the probability of meeting the expected end date can be calculated [12]. Kelley [22] defines the main goal of PERT as monitoring progress of projects, which is also in the letters E and R of PERT.

The Critical-Path Method is based on a linear program that computes the utility of a project as function of its duration. The optimal schedule is defined as the schedule with the maximum utility among all feasible schedules with the same project duration [22]. Though CPM can provide optimal schedules, it can not be used for scheduling clinical pathways. If the critical path, i.e. the patient with the longest length of stay, is scheduled first it would be unethical.

Enterprise Resource Planning (ERP) is another scheduling technique frequently used in industries. ERP resulted from Material Requirements Planning-II with the goal to provide an overview of entire enterprises by sorting data and obtaining feasible schedules for all departments of the enterprise, taking capacity constraints into account [20]. Although ERP could be the solution to scheduling CPs, it can not be applied for determining the effect of reservation on the blocking probability since the probability depends on the instances the ERP programme receives. Moreover ERP-software is proved to be non-flexible, complex, difficult to implement and little compatible with existing software [35].

Treatment plans of rehabilitation patients consist of a set of departments and a number of treatments for each department. Scheduling clinical pathways could therefore be seen as an interval scheduling problem, where one patient requires multiple intervals at multiple servers. Scheduling CPs is then equivalent to an independent set problem, since scheduled intervals must be non-overlapping. In [18] Halldórsson gives a literature review on algorithms approximating independent sets in interval graphs. Because all algorithms investigated in [18] approximate the independent set, it is hard to investigate the impact of reservation on the exact blocking probability with these algorithms.

There is many interesting literature on scheduling intervals, for example see [6], [16], [17] and [27]. Arkin and Silverberg ([2]) describe an algorithm that turns an interval scheduling problem into determining the longest path. Although this algorithm can provide an exact set of accepted intervals, the algorithm becomes rather complicated for scheduling multiple intervals per patient and multiple CPs at the same time. Moreover actually scheduling intervals and comparing them with the schedules without reservation, seems to depend heavily on the given instance and it is therefore hard to say anything in general. This holds for all literature on interval scheduling problems.

Concluding it can be said that the most frequently used scheduling techniques from the industries can not be used to schedule clinical pathways and/or can not be used to evaluate the impact of scheduling with CPs. Therefore we only focus on finding a way to analyze the impact. Here queuing theory appears to be a promising option, as will be clarified below.

2.2 Queuing theory

For this research we looked at literature on queueing models that could be usefull for analyzing the impact of reservation on the blocking probability. The first article worth mentioning is written by Bonald and considers non-Poisson arrivals to an Erlang model [5]. Here Bonald introduces *sessions*, which represent multiple calls from one customer (Bonald applies his model to a telephone network). In [5] it is claimed for the insensitivity properties to hold, it is sufficient that users generate a session according to a Poisson process, each session being composed of a random number of calls and idle periods. Each call of the session can be blocked and will then be lost. If the blocked call was the last of the session, the user will leave the system and otherwise the user will wait a random time untill the next call of the session requires service.

The major result in [5] is that all calls (treatments) of the session have the same blocking probability, which implies that each treatment independently could be blocked. Although the results of this article are analytic and given with extensive proofs, they can not be applied to the network of clinical pathways since we do not allow a treatment of a patient being blocked once he/she has already received some treatments.

Another article worth mentioning is written by Berezner and Krzesinski [4]. They examine a circuit-switched network with C source switches all linked to one destination switch through one tandem switch (see Figure 2.1). In this system calls can queue at different switches. The queues hold some of the blocked calls and these calls are served First-Come-First-Serve when there is a

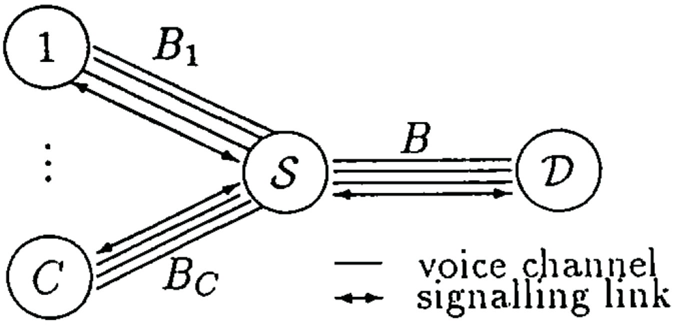


Figure 2.1: Queuing model described in [4].

circuit available. The authors propose different places of the queue in the network: at the sources or the tandem switch. They claim that the queues make the blocking probability decline at the expense of a small connection delay. However, the article contains litte exact results and relies on an algorithm to calculate the blocking probability.

In [3] a similar queueing model is described. Here customers (arriving randomly) are served by a queueing system consisting of a sequence of two service stations with infinite queue allowable before the first station and no queue allowable between the stations. We could model the reservation time as the service time in the first queue, taking infinitely many servers at the first station. In the tandem queue considered in [3] customers can be blocked when their service at the first station is completed at a time instance the second station is occupied. In this sense there was no advance capacity reservation and only unnecessary waiting of patients that are not served.

Roberts and Liao ([29]) analyze a traffic model where part of the arriving phonecalls reserve capacity in advance. The authors try to model the system with a queueing model, but fail to derive analytic results. Their second approach, to model the reservation traffic process with a birth process, results in simple closed form expressions that enable the authors to estimate the required number of channels to carry the offered traffic load with low blocking probability. In this research we model the reservation model with a queueing model and obtain analytic results of it.

In [11] the authors consider a “two-stage no-wait tandem queueing system, in which any customer who finds all servers busy at his destination stage will be lost”. An admission control is applied such that all customers will go through the second stage successfully. The authors compare the loss rate of the controlled tandem queue with the loss rate in the uncontrolled tandem queue. Compared to our reservation model, we could model the reservation time of the patients as the time in the first stage of the tandem queue described in [11], assuming this first stage has infinitely many servers. However, Chang and Chen give in their article only numerical results and their main result is that both loss rates are equal when there is only one server at the first stage.

Harms and Wong ([19]) consider a loss system where the service center rejects a request if it can not accommodate the request at the preferred start time. Each request is characterized by a start time, a bandwidth requirement and a holding time. In this article the authors investigate the benefits of delaying scheduling decisions from the performance point of view, which would be an undesirable option for scheduling patients.

In [38] an optimal access strategy for managing reservation systems with Poisson arrivals is studied. Virtamo and Aalto consider a reservation book in which arrivals can occupy several timeslots and they describe an optimal strategy for accepting arrivals with the aim to maximize the expected system utilization. However, when these arrivals are patients this selective acceptance

would be unethical.
In [37] Virtamo considers the reservation book without an acceptance strategy. Here he models the state of the network as "islands of free time-slots", with the islands representing an uninterrupted series of free time-slots. For this state description to be valid, Virtamo has to assume that the starting times (the time each service has to start) are uniformly distributed. Although the results of [37] would probably still hold for exponential arrivals, the state description would become very complicated when one patient has to receive treatment from multiple servers.

The queuing model that seems most suitable for modelling a network of clinical pathways is a loss network with fixed routing. Ross ([30]) elaborates this queuing model for a telephone network with arriving calls requiring a route via certain links. In the network there is a certain number of circuits between adjacent links. If an arriving call finds one of the links it requests occupied, it is blocked and lost.
In the mapping from the telephone to the patient network, calls are mapped to patients, routes to clinical pathways, links to departments and the number of circuits between links is denoted by the capacity of each department. The main difference between telecommunication networks and networks of clinical pathways, is the way the resources are confiscated. When a call arrives in a telecom network and the required circuits are available, all necessary resources are claimed instantaneously. Arriving patients require more than one appointment and can therefore claim multiple resources in the future. Moreover arriving patients that find insufficient recourses available are not blocked instantaneously, since they are allowed to wait until their required resources are available (but only within a specified norm).

De Bruin et al. ([14]) also use loss networks in a health care research. In this article a decision support system for evaluating the current size of nursing units is developed. This decision support tool is based on the Erlang loss model.

A review of literature on the extended loss network (with reservation of resources) does not provide many articles. Almost twenty years after the famous article on loss networks by Kelly in 1991 [23], Zachary and Ziedins also wrote an article on the state of the art of loss networks [39]. Zachary and Ziedins do not mention reservation in loss networks or a concept which resembles reservation.

In an article by Lu and Radovanović [25] the possibility of reservation of resources in loss networks is introduced. The authors do so for a network in which arrivals follow a compound renewal process and have random resource requirements. (They model these requirements as subexponential random variables.) One of the results in this article is that the stationary blocking probability for a single resource network without reservation is approximately equal to the tail of the resource requirement distribution. Perhaps the most interesting result is that they show that the blocking probability in the system with advance reservations approaches the blocking probability in the system without reservation when the capacity grows to infinity. However the assumption of the capacity going to infinity is not realistic for rehabilitation centres, the authors give this result without a formal proof and do not give error bounds for limited capacity.

From this we can conclude that the theory of loss networks needs to be extended before it is applicable to networks of clinical pathways.

Chapter 3

Model description

In this section the mathematical model of the network of clinical pathways will be clarified. Here the notation that will be used throughout the report is introduced as well as the characteristics of the network.

The network of CPs with the possibility of reservation is modelled as an inifinite server queue ($M|G|\infty$) in tandem with a loss queue with capacity C ($M|G|C|C$). Patients arrive to the first queue according to a Poisson process (with parameter λ) and declare a remaining service time at both queues instantaneously. As a consequence the time this patient leaves the system is known at arrival.
When we model the network in this way, we assume each patient has a random reservation time (this is the time in the $M|G|\infty$ queue) before entering service (time in the $M|G|C|C$ queue). If at arrival the capacity at the second queue is insufficient in the time interval the patient requires it, he/she is blocked and leaves the system.

We denote the state of the network by $(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i))$ when there are i patients in the network at time t, with $N_{R,S}$ the number of patients in reservation (first queue) and service (second queue), r_k the remaining reservation time (first queue) and s_k the remaining service time (second queue) for $k = 1, ..., i$.

In words the possible transitions for this network are:

- An arrival to the network,
- A service at the first queue is completed,
- A service at the second queue is completed (patient leaves the network).

If there is no transition in a small time interval then either r_k or s_k decreases, because either the patient is at the first or the second queue. The first transition can only occur if the new patient (with characteristic (r_{i+1}, s_{i+1})) does not violate the capacity constraint. Before writing down the blocking rule, we first define a function $N\left(t|\hat{t}\right)$ that equals the number of patients present at the second queue at time t given the state of the network at time $\hat{t}$. When the state of the system is $(N_R, N_S, (r_1, s_1), ..., (r_i, s_i))$ at time $\hat{t}$, $N\left(t|\hat{t}\right)$ is defined by:

$$N\left(t|\hat{t}\right) = \sum_{k=1}^{i} \mathbb{1}_{\{t \in [r_j, r_j + s_j]\}} ,$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator of the event $\{\cdot\}$. An $(i+1)^{th}$ patient arriving at time $\hat{t}$ is only accepted if $N\left(t|\hat{t}\right) < C$ at any time $t \in \left[\hat{t} + r_{i+1}, \hat{t} + r_{i+1} + s_{i+1}\right]$, and the check that needs to be performed can be written as:

$$N\left(t|\hat{t}\right) \leq C-1 \text{ for } t \in [\hat{t}+r_{i+1}, \hat{t}+r_{i+1}+s_{i+1}] \ . \tag{3.1}$$

Here it is checked that in the interval $[\hat{t}+r_{i+1}, \hat{t}+r_{i+1}+s_{i+1}]$ (in which the arriving patient is at the second queue) there is at least one unit of capacity available. Notice that with this definition of the state space, the value of $N\left(t|\hat{t}\right)$ is known in advance on the interval $\left[\hat{t}, \hat{t}+\max\limits_{1\leq j\leq i}(r_j+s_j)\right]$.

Now we can denote the state space of this tandem queue by:

$$S = \Big\{ N_R, N_S, (r_1, s_1), ..., (r_{N_R+N_S}, s_{N_R+N_S}) | N_S \leq C, s_k \in \mathbb{R}^+/\{0\},$$
$$r_k \in \mathbb{R}^+, N_R, N_S \in \mathbb{N}^+, \forall t \in [0, \max_k(r_k+s_k)] : N(t) \leq C \Big\}$$

With this information we have defined our queueing model representing a loss network with reservation. Ideally we want to have the stationary distribution of our tandem queue, since with this distribution we can derive important performance measures like the probability an arriving patient is blocked. Finding a stationary distribution simplifies a lot if the distribution can be written in a closed form. The next two sections however, show that finding this stationary distribution is hard and that the reservation model does not have a product-form solution.

Chapter 4

The infinite-server case

In this chapter we first formulate and secondly solve the Kolmogorov differential equations for the infinite-server reservation model. Recently Taylor ([34]) was the first to formally state and prove the stationary distribution of a loss network where the elapsed service time is in the state description. In [34] the stationary distribution of a loss queue is obtained by formulating the Kolmogorov differential equations and showing that the proposed stationary distribution is a solution of the equations. Since we do not know the stationary distribution of the reservation model in advance (like Taylor did before he wrote his article), the stationary distribution of the infinite-server reservation model is solved first. This stationary distribution must resemble the distribution of the ordinary tandem of infinite server queues found by Boxma ([8]).
Often for loss queues, truncating the infinite-server distribution results in the desired stationary distribution. In the following section we expand the proof of Taylor to the case of the infinite-server reservation model. In our research we were unable to find the stationary distribution of the reservation model by solving the differential equations. In the last section of this chapter we indicate why we thought we could derive the stationary distribution of the reservation model by truncating the infinite server case. In the next chapter we derive the stationary distribution by means of a renewal theory argument.

4.1 Differential equations

As said before we investigate the tandem of two $M|G|\infty$ queues where the service requirements for both queues are drawn at arrival. In this section we obtain differential equations for the stationary distribution by conditioning on a small time Δ ago. In order to do this we define the reservation time of a patient with random variable R and the time in the service queue with random variable S. We assume R and S are independently drawn when the patient enters the rehabilitation centre. We define $B_R\,(t)$ and $B_S\,(t)$ the cumulative distribution functions of R and S. Therefore we define $b_S(y) = \frac{dB_S}{dy}(y)$ and $b_R(y) = \frac{dB_R}{dy}(y)$.

If we let $\pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i))$ be the probability the system is in state $(N_R, N_S, ...$ $..., (r_1, s_1), ..., (r_i, s_i))$ at time t with $N_R + N_S = i$, then the stationary distribution is defined by $\lim_{t\to\infty} \pi(t : N_R, N_S, (r_1, s_1), ...\ ..., (r_i, s_i))$. W.l.o.g. we assume that the patients in the statevector are ordered such that the first N_R patients are at the first queue and the last N_S at the second. For defining the differential equations, we can split up two different cases: $\{N_R = 0, N_S = 0\}$ and the case where at least one queue contains one patient. For $\{N_R = 0, N_S = 0\}$ it holds:

$$\begin{aligned}\pi\{t : N_R = 0, N_S = 0\} =&(1 - \lambda\Delta)\pi\{t - \Delta : N_R = 0, N_S = 0\}\\ &+ \int_0^\Delta \pi(t - \tau : N_R = 0, N_S = 1, (0, \tau))\,\mathrm{d}\tau,\end{aligned}$$

because either nothing happend or a service at the second queue is completed in time interval Δ and there was nobody left in the system.

For $N_R > 0$ and/or $N_S > 0$, $r_k > \Delta$ for $1 \leq k \leq N_S$ and $s_k > \Delta$ for $N_S < k \leq N_S + N_R$:

$$\begin{aligned}&\pi(t : N_R = i - m, N_S = m, (r_1, s_1), ..., (r_i, s_i)) =\\
&\quad \frac{\lambda\Delta}{i-m}\sum_{k=1}^{i-m} \pi\left(t - \Delta : N_R = i - m - 1, N_S = m, (r_1 + \Delta, s_1), ...\right.\\
&\qquad\quad \left. ..., (r_{k-1} + \Delta, s_{k-1}), (r_{k+1} + \Delta, s_{k+1}), ..., (r_i, s_i + \Delta)\right) b_R(r_k) b_S(s_k)\\
&\quad + \sum_{k=i-m+1}^{i} \int_0^\Delta \pi\left(t - \tau : N_R = i - m, N_S = m + 1, (r_1 + \tau, s_1), ...\right.\\
&\qquad\quad \left. ..., (r_k, s_k + \tau), (0, \tau), (r_{k+1}, s_{k+1} + \tau), ..., (r_i, s_i + \tau)\right) \mathrm{d}\tau\\
&\quad + \frac{1}{m}\sum_{g=i-m}^{i}\sum_{j=1}^{m} \int_0^\Delta \pi\left(t - \tau : N_R = i - m + 1, N_S = m - 1, (r_1 + \tau, s_1), ..\right.\\
&\qquad\quad \left. ..., (r_{j-1} + \tau, s_{j-1}), (\tau, s_g), (r_j + \tau, s_j), .., (r_i, s_i + \tau)\right) \mathrm{d}\tau\\
&\quad + (1 - \lambda\Delta)\, \pi\left(t - \Delta : N_R = i - m, N_S = m, (r_1 + \Delta, s_1), ...\right.\\
&\qquad\quad \left. ..., \left(r_{i-m} + \Delta, s_{i-m}\right), \left(r_{i-m+1}, s_{i-m+1} + \Delta\right), .., (r_i, s_i + \Delta)\right)\\
&\quad + o\left(\Delta\right).\end{aligned}$$

Here the first expression on the right hand side represents the event that the k^{th} patient has just arrived. The $\frac{1}{i-m}$ is in this expression because the patient has $i-m$ possible locations in the state vector to go to. The second term on the right hand side represents that there was a departure at the second queue. The third expression represents the events that one of the m patients at the second queue (j) arrived there in time interval Δ from one of the $(i - m + 1)$ patients that were at the reservation queue a time Δ ago. The $\frac{1}{m}$ is in this expression because the patient has m possible locations in the state vector to go to. The last two expressions stand for the events that that no transitions occurred during time interval Δ or that two events have occurred. Obviously the fractions in the first and third term on the right hand side are only valid when $i - m > 0$ and $m > 0$.

With these equations we can obtain differential equations by dividing by Δ and letting $\Delta \to 0$. For $\{N_S = 0, N_R = 0\}$:

$$\begin{aligned}\frac{d}{dt}\pi(t : N_S = 0, N_R = 0) =&\lambda\pi(t : N_S = 0, N_R = 0)\\ &+ \pi(t : N_S = 1, N_R = 0, (0, 0^+)),\end{aligned}$$

where 0^+ denotes the limit to 0 from above. And since we know this time-derivative must equal zero (otherwise the solution would not be stationary), we obtain:

$$\pi(N_S = 0, N_R = 0) = \frac{1}{\lambda}\pi(N_S = 1, N_R = 0, (0, 0^+)).$$

These pobabilities do not depend on time and therefore the variable t is left out.

For $N_R > 0$ and/or $N_S > 0$, $r_k > \Delta$ for $1 \le k \le N_S$ and $s_k > \Delta$ for $N_S < k \le N_S + N_R$ we obtain in the same way:

$$\begin{aligned}
&\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s_1} - ... - \frac{\partial}{\partial r_i}\right) \pi(t : N_R = i-m, N_S = m, (r_1, s_1), ..., (r_i, s_i)) = \\
&\quad \frac{\lambda}{i-m} \sum_{k=1}^{i-m} \pi\left(t : N_R = i-m-1, N_S = m, (r_1, s_1), ...\right. \\
&\qquad \left. ..., (r_{k-1}, s_{k-1}), (r_{k+1}, s_{k+1}), ..., (r_i, s_i)\right) b_R(r_k) b_S(s_k) \\
&\quad + \sum_{k=i-m+1}^{i} \pi\left(t : N_R = i-m, N_S = m+1, (r_1, s_1), ...\right. \\
&\qquad \left. ..., (r_k, s_k), (0, 0^+), (r_{k+1}, s_{k+1}), ..., (r_i, s_i)\right) \\
&\quad + \frac{1}{m} \sum_{g=i-m}^{i} \sum_{j=1}^{m} \pi\left(t : N_R = i-m+1, N_S = m-1, (r_1, s_1), ..\right. \\
&\qquad \left. ..., (r_{j-1}, s_{j-1}), (0^+, s_g), (r_j, s_j), .., (r_i, s_i)\right) \\
&\quad - \lambda \pi(t : N_R = i-m, N_S = m, (r_1, s_1), ..., (r_i, s_i)).
\end{aligned}$$

Again, we know that the time derivative must equal zero. Therefore the stationary distribution must be a solution to the following equations:

$$\begin{aligned}
&\left(- \frac{\partial}{\partial s_1} - ... - \frac{\partial}{\partial r_i}\right) \pi\left(N_R = i-m, N_S = m, (r_1, s_1), ..., (r_i, s_i)\right) = \\
&\quad \frac{\lambda}{i-m} \sum_{k=1}^{i-m} \pi\left(N_R = i-m-1, N_S = m, (r_1, s_1), ...\right. \\
&\qquad \left. ..., (r_{k-1}, s_{k-1}), (r_{k+1}, s_{k+1}), ..., (r_i, s_i)\right) b_R(r_k) b_S(s_k) \\
&\quad + \sum_{k=i-m+1}^{i} \pi\left(N_R = i-m, N_S = m+1, (r_1, s_1), ...\right. \\
&\qquad \left. ..., (r_k, s_k), (0, 0^+), (r_{k+1}, s_{k+1}), ..., (r_i, s_i)\right) \\
&\quad + \frac{1}{m} \sum_{g=i-m}^{i} \sum_{j=1}^{m} \pi\left(N_R = i-m+1, N_S = m-1, (r_1, s_1), ...\right. \\
&\qquad \left. ..., (r_{j-1}, s_{j-1}), (0^+, s_g), (r_j, s_j), .., (r_i, s_i)\right) \\
&\quad - \lambda \pi(N_R = i-m, N_S = m, (r_1, s_1), ..., (r_i, s_i)).
\end{aligned}$$

In the next section we will give the solution to these differential equations and since the stationary distribution is unique, this implies we have found the stationary distribution for the infinite server network.

4.2 Solution

Now we have obtained the differential equations the stationary distribution must satisfy, we can prove the following theorem:

Theorem 1. *For the tandem network $M|G|\infty \rightarrow M|G|\infty$, the stationary distribution is given by:*

$$\pi\left(N_R = i-m, N_S = m, (r_1,s_1),...,(r_i,s_i)\right) = \frac{(\lambda E(R))^{i-m}}{(i-m)!}\frac{(\lambda E(S))^m}{m!}G^{-1}\prod_{k=1}^{i} h\left(r_k,s_k\right) \quad (4.1)$$

Here G^{-1} is a normalization constant and

$$h(r_k,s_k) = \frac{\partial}{\partial r_k \partial s_k} P\left\{R_k^{res} \leq r_k, S_k^{res} \leq s_k\right\},$$

with R_k^{res}, S_k^{res} the residual service time of patient k at the first and second queue.

This solution is of product-form, since it is the product of the stationary distributions of the infinite server queues when seen in isolation.
We derive $h(r_k,s_k)$ directly by conditioning on which queue the patient is in:

$$h(x,y) = \begin{cases} b_S(y)\frac{1-B_R(x)}{E(R)} & \text{if } x>0 \\ \frac{1-B_S(y)}{E(S)} & \text{if } x=0 \end{cases}$$

This discontinuous function of h follows from the fact that if $x > 0$ the residual service time in the second queue can not be different from the service time at arrival. When $x > 0$ the cumulative distribution function of the residual service time in the first queue is given by:

$$\frac{\int_0^x [1-B_R(u)]\,\mathrm{d}u}{E(R)}.$$

And when we take the derivative of this to x, we obtain the expression

$$\frac{\partial}{\partial x}P(R^{res} \leq x) = \frac{1-B_R(x)}{E(R)}.$$

In the same way, when we know that if $x = 0$, the probability that the remaining service time at the second queue is approximately y is $\frac{1-B_S(y)}{E(S)}$.
We know h is not a probability denstity function. However in this text we write h without the probability of $\{x = 0\}$ in order to make the functions less complicated. This is allowed since these probabilities can be contained within the normalization constant G^{-1}. In the steps of the proof below, these probabilities occur on both the left and right hand side of the equality sign and therefore it has no impact if we leave them out.

We now prove that the proposed stationary distribution solves the differential equations given above.

Proof. The first thing to notice is that the left hand side of the differential equations consists of two types of derivatives. In the limiting distribution the only term that depends on r_k or s_k is $h(r_k,s_k)$. Therefore the two types of partial derivatives of h are given by:

$$\begin{aligned}\frac{\partial h(r_k,s_k)}{\partial s_k} &= \frac{\partial}{\partial s_k}\left[\frac{1-B_S(s_k)}{E(S)}\right] \\ &= -\frac{b_S(s_k)}{E(S)},\end{aligned}$$

since for this case we know that $r_k = 0$. Therefore we obtain for $N_R < k \leq (N_R + N_S)$:

$$\begin{aligned}
&\frac{\partial \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i))}{\partial s_k} \\
&\quad = -\frac{b_S(s_k)}{E(S)h(r_k, s_k)} \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i)) \\
&\quad = -\frac{b_S(s_k)}{E(S)} \frac{E(S)}{1 - B_S(s_k)} \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i)) \\
&\quad = -\frac{b_S(s_k)}{1 - B_S(s_k)} \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i)).
\end{aligned}$$

In the same way we can derive the partial derivatives for $1 \leq k < N_R$:

$$\begin{aligned}
\frac{\partial h(r_k, s_k)}{\partial r_k} &= \frac{\partial}{\partial r_k}\left[b_S(s_k)\frac{1 - B_R(r_k)}{E(R)}\right] \\
&= -\frac{b_S(s_k)b_R(r_k)}{E(R)},
\end{aligned}$$

since for this case we know that $r_k > 0$. Therefore we can say for $1 \leq k < N_R$:

$$\begin{aligned}
&\frac{\partial \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i))}{\partial r_k} \\
&\quad = -\frac{b_S(s_k)b_R(r_k)}{E(R)h(r_k, s_k)} \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i)) \\
&\quad = -\frac{b_S(s_k)b_R(r_k)}{E(R)} \frac{E(R)}{(1 - B_R(r_k))b_S(s_k)} \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i)) \\
&\quad = -\frac{b_R(r_k)}{1 - B_R(r_k)} \pi(t : N_R, N_S, (r_1, s_1), ..., (r_i, s_i)).
\end{aligned}$$

From this it follows that we can write the differential equations as:

$$\begin{aligned}
&-\left(\sum_{k=1}^{i-m} -\frac{b_R(r_k)}{1 - B_R(r_k)} + \sum_{k=i-m+1}^{i} -\frac{b_S(s_k)}{1 - B_S(s_k)}\right) \pi(t : N_R = i - m, N_S = m, (r_1, s_1), ..., (r_i, s_i)) = \\
&\quad \frac{\lambda}{i - m} \sum_{k=1}^{i-m} \pi\, (N_R = i - m - 1, N_S = m, (r_1, s_1), ... \\
&\qquad ..., (r_{k-1}, s_{k-1}), (r_{k+1}, s_{k+1}), ..., (r_i, s_i))\, b_R(r_k)b_S(s_k) \\
&\quad + \sum_{k=i-m+1}^{i} \pi \left(N_R = i - m, N_S = m + 1, (r_1, s_1), ..., (r_k, s_k), (0, 0^+), (r_{k+1}, s_{k+1}), ..., (r_i, s_i)\right) \\
&\quad + \frac{1}{m} \sum_{j=i-m}^{i} \sum_{g=1}^{m} \pi\, (N_R = i - m + 1, N_S = m - 1, (r_1, s_1), ... \\
&\qquad ..., (r_{j-1}, s_{j-1}), (0^+, s_g), (r_j, s_j), .., (r_i, s_i)\Big) \\
&\quad - \lambda\pi(N_R = i - m, N_S = m, (r_1, s_1), ..., (r_i, s_i)).
\end{aligned}$$

And if we express the stationary probabilities in $\pi(t : N_R = i-m, N_S = m, (r_1, s_1), ..., (r_i, s_i))$ we obtain:

$$\left(\sum_{k=1}^{i-m}\frac{b_R(r_k)}{1-B_R(r_k)}+\sum_{k=i-m+1}^{i}\frac{b_S(s_k)}{1-B_S(s_k)}\right)\pi(t:N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i))=$$

$$+\frac{\lambda}{i-m}\sum_{k=1}^{i-m}\frac{\pi(N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i))}{h(r_k,s_k)}\frac{i-m}{\lambda E(R)}b_R(r_k)b_S(s_k)$$

$$+\sum_{k=i-m+1}^{i}\pi(N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i))\frac{\lambda E(S)}{m+1}h(0,0^+)$$

$$+\frac{1}{m}\sum_{j=i-m}^{i}\sum_{g=1}^{m}\pi\left(N_R=i-m,N_S=m,(r_1,s_1),...\right.$$

$$\left....,(r_i,s_i)\right)\frac{m}{\lambda E(S)}\frac{\lambda E(R)}{i-m+1}\frac{h(0^+,s_j)}{h(0,s_j)}$$

$$-\lambda\pi(N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i)).$$

And if we fill in the values of h and simplify the functions, we obtain:

$$\left(\sum_{k=1}^{i-m}\frac{b_R(r_k)}{1-B_R(r_k)}+\sum_{k=i-m+1}^{i}\frac{b_S(s_k)}{1-B_S(s_k)}\right)\pi(t:N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i))=$$

$$+\sum_{k=1}^{i-m}\frac{b_R(r_k)}{1-B_R}\pi(N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i))$$

$$+\lambda\pi(N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i))$$

$$+\sum_{j=i-m}^{i}\frac{b_S(s_j)}{1-B_S}\pi(N_R=i-m,N_S=m,(r_1,s_1),..,(r_i,s_i))$$

$$-\lambda\pi(N_R=i-m,N_S=m,(r_1,s_1),...,(r_i,s_i)).$$

Now we see that the second and the last expression on the right hand side cancel and then the terms on the left and right hand side of the equality sign say exactly the same. With this result it is shown that the stationary distribution (4.1) holds for the case where there is no capacity constraint. □

4.3 Truncation for the finite capacity case

In the previous section we proved (4.1) is the stationary distribution of two infinite-server queues in tandem. In this section we point out how we can truncate the infinite server case of an ordinary tandem in a way the product-form of the solution is preserved.

In Figure 4.1 the possible transitions for two queues in an ordinary tandem are given. Here each pair (x,y) represents x patients at the first queue and y at the second. In the reservation model the pairs (x,y) represent a group of states and the arrival rate λ is different since it is state dependent in the reservation model. When each queue of the ordinary tandem of infinite server queues is seen in isolation, the partial balance property holds. Partial balance holds when the limiting distribution π satisfies for each $j=0,1,2$:

$$\sum_{k=0}^{2}\pi(\mathbf{n})q(\mathbf{n},\mathbf{n}-\mathbf{e_j}+\mathbf{e_k})=\sum_{k=0}^{2}\pi(\mathbf{n}-\mathbf{e_j}+\mathbf{e_k})q(\mathbf{n}-\mathbf{e_j}+\mathbf{e_k},\mathbf{n}),$$

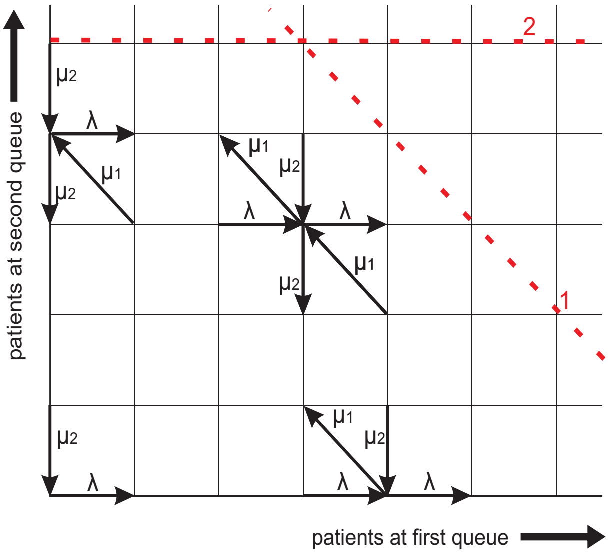


Figure 4.1: Transition graph of tandem of two queues.

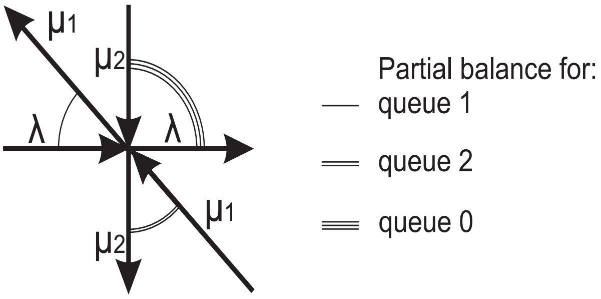


Figure 4.2: Partial balance in the transition graph.

with state $\mathbf{n} = (n_1, n_2)$ a vector containing the number of patients at each queue, $q(\mathbf{n}, \mathbf{n} - \mathbf{e_j} + \mathbf{e_k})$ the transition rate from state $\mathbf{n}$ to state $\mathbf{n} - \mathbf{e_j} + \mathbf{e_k}$ (thus a transition from queue j into queue k). Here transitions from and to queue 0 represents arrivals to and departures from the tandem. In Figure 4.2 it is pointed out which probability flows (i.e. $\pi(\mathbf{n})q(\mathbf{n}, \mathbf{n} - \mathbf{e_j} + \mathbf{e_k})$) are balanced if the partial balance equations hold.

When the tandem queue is truncated according to red dotted line 1 (for example $n_1 + n_2 \leq C$ for some integer C), the stationary distribution of the tandem queue is of product form since two balanced transitions are cancelled. Therefore we say (in accordance with [26]) the set of states that remains after truncation has the same stationary distribution as the infinite server case (up to a normalization), which is of product form.
As can be seen in Figures 4.1 and 4.2 a truncation along the second red dotted line does not result in a product form stationary distribution, since the transitions that are cut off are not balanced. This truncation denotes that only the second queue has capacity constraint $n_2 \leq C$. In [7] it is shown that in this case for example the STOP-protocol or Jump-over blocking has to be applied in order to preserve the product form stationary distribution. The STOP-protocol prescribes that once the second queue is saturated, service at all other stations is stopped and arrivals to the network are discarded. Under jump-over blocking, a patient that is blocked at the second queue jumps over this station at infinite speed and attempts to enter a next queue (in the reservation

model this implies leaving the system) as if it was served normally.
In the reservation model a different kind of alteration is applied: here patients can never be served at the first queue when the second queue is saturated in the time interval the patient requires service at the second queue. In other words, a patient that requires service at an instace in which there is insufficient capacity at the second queue, will be blocked at the first (reservation) queue and therefore never access the system. The result of this blocking rule resembles the result of the STOP-protocol since we enforce that when the second queue is saturated, there will always be a departure from the second queue before a departure from the first queue occurs. Arrivals can be accepted, but only if they do not require service at the second queue when its saturated. In this way all patients that enter the system will never be blocked when entering the second queue.
With this reasoning we thought partial balance would hold for the state space of the reservation model, but we were unable to succesfully derive a product form stationary distribution for this network. We therefore derived the stationary distribution of a simplified reservation model by means of renewal theory. This is clarified in the next chapter. It turns out that the stationary distribution does not have product-form, which was somewhat surprising.

Chapter 5

The deterministic single-server case

In this chapter we will obtain the stationary distribution in a direct way, with an argument of renewal theory (see for example [31]). For a start we look at a reservation model with deterministic reservation time D_2 and deterministic service time D_1. We assume that $D_2 < D_1$, which enforces there can be at most two patients in the system at the same time. Below a possible realisation of this queuing network is given.

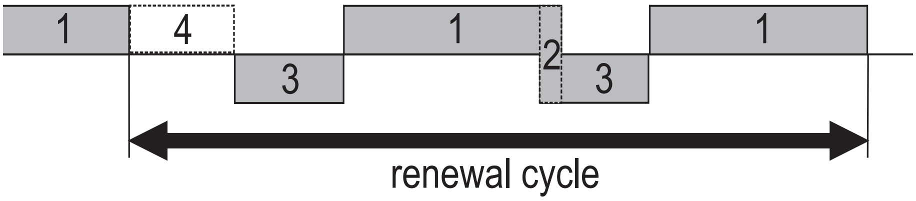


Figure 5.1: Realisation of the reservation model.

In Figure 5.1 a patient in the $M|D|1|1$ queue is represented by a block above the (hotizontal) time axis and a patient in reservation is represented by a block below the time axis. Here we choose to define that a renewal occurs when a departing patient leaves the system empty. As can be seen in Figure 5.1, we make a distinction between the following four events:

1. $\{N_R = 0, N_S = 1\}$,
2. $\{N_R = 1, N_S = 1\}$,
3. $\{N_R = 1, N_S = 0\}$,
4. $\{N_R = 0, N_S = 0\}$.

For each event we obtain the (time independent) cumulative density function by calculating the expected time this event occurs per cycle and devide this by the expected cycle length. The density functions of the four events will be derived in seperate sections. The concluding section consists of a list of all four density functions.

5.1 The event $\{N_R = 0, N_S = 1\}$

In order to obtain $P\{N_R = 0, N_S = 1, S^{res} \geq x\}$, we will obtain an expression for the expected time per cycle the event $\{N_R = 0, N_S = 1\}$ occurs, with S^{res} the elapsed service time and $0 \leq x < D_1$. As can be seen in Figure 5.2 two situations can occur in one arbitrairy service time:

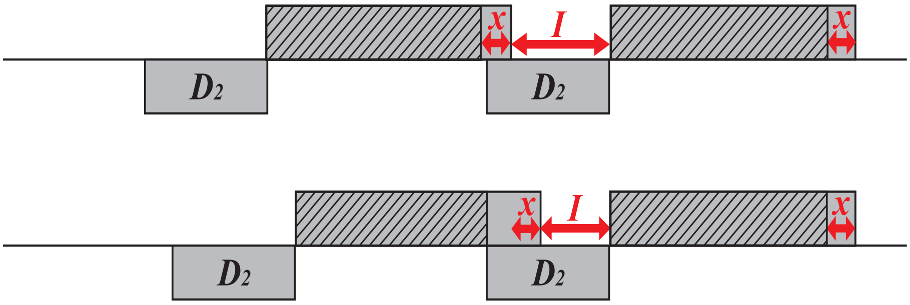


Figure 5.2: Possibilities for the length of interval $\{N_R = 0, N_S = 1\}$.

- $I + x > D_2$, then the length of the event $\{N_R = 0, N_S = 1\}$ is $D_1 - x$,
- $I + x \leq D_2$, then the length of the event $\{N_R = 0, N_S = 1\}$ is $D_1 - (D_2 - I)$.

From this we can calculate for each arbitrary service time in the cycle:

$$
\begin{aligned}
&E[\text{event } \{N_R = 0, N_S = 1\} \text{ per patient}] \\
&\quad = \int\limits_{0}^{D_2-x} (D_1 - D_2 + i)\lambda e^{-\lambda i} di + \int\limits_{D_2-x}^{\infty} (D_1 - x)\lambda e^{-\lambda i}\,\mathrm{d}i \\
&\quad = -(D_1 - D_2 + i)e^{-\lambda i}\Big|_{i=0}^{D_2-x} - \int\limits_{0}^{D_2-x} -e^{-\lambda i}\,\mathrm{d}i + (D_1 - x)(e^{-\lambda(D_2-x)} - 0) \\
&\quad = D_1 - D_2 - \frac{1}{\lambda}(e^{-\lambda(D_2-x)} - 1).
\end{aligned}
$$

Here the first term on the right hand side occurs by conditioning on the interarrival time I, where I is exponentially distributed. The second term is by partial integration of the first.

We can also see that the cycle length has a geometric distribution, since the cycle ends if there is one interarrival time greater than D_2 and continues otherwise. The probability a cycle ends is $p = e^{-\lambda D_2}$, since this tis the probability $P(I > D_2)$ with I exponentially distributed. Therefore the average number of patients treated in one cycle is $\frac{1}{p} = e^{\lambda D_2}$. Per patient the average treatment time (without reservation!) is $D_1 + \frac{1}{\lambda}$, so we can state the average cycle length is $(D_1 + \frac{1}{\lambda})e^{\lambda D_2}$. From this we can conclude:

$$
P\{N_R = 0, N_S = 1, S^{res} \geq x\} = \frac{[D_1 - D2 - \frac{1}{\lambda}(e^{-\lambda(D_2-x)} - 1)]e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda})e^{\lambda D_2}}.
$$

5.2 The event $\{N_R = 1, N_S = 1\}$

In this section we obtain an expression for $P\{N_R = 1, N_S = 1, S^{res} \geq x, R^{res} \leq y\}$, with R^{res} the elapsed reservation time, $0 \leq x < D_2$ and $0 \leq y < D_2$. In Figure 5.3 we can see that four situations can occur in one arbitrary service time:

- $y < I + x \leq D_2$, then the event $\{N_R = 1, N_S = 1\}$ has length $D_2 - (I + x)$,
- $I + x \leq y \leq D_2$, then the event $\{N_R = 1, N_S = 1\}$ has length $D_2 - y$,
- $y < I$, then the event $\{N_R = 1, N_S = 1\}$ does not occur,

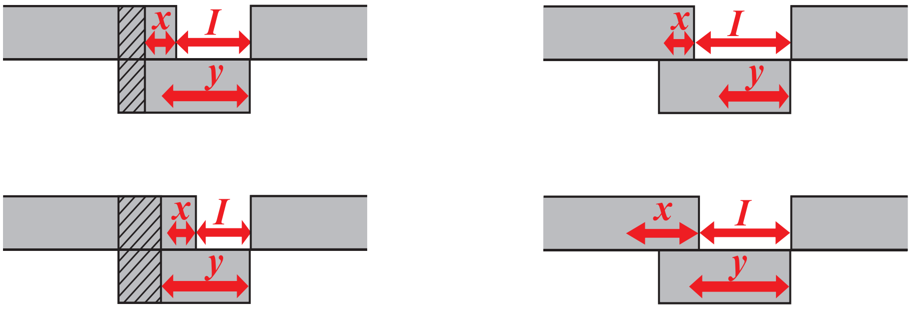


Figure 5.3: Possibilities for the length of interval $\{N_R = 1, N_S = 1\}$.

- $I + x > D_2$, then the event $\{N_R = 1, N_S = 1\}$ does not occur.

By conditioning on the length of the interarrival time, we obtain:

$$E[\text{event } \{N_R = 1, N_S = 1\} \text{ per patient}]$$
$$= \int_0^{(y-x)^+} (D_2 - y)\lambda e^{-\lambda i}\, \mathrm{d}i + \int_{(y-x)^+}^{D_2-x} (D_2 - i - x)\lambda e^{-\lambda i}\, \mathrm{d}i.$$

And if we split out the two cases $y \geq x$ and $y < x$, we can calculate the expected interval length. For $y \geq x$ these expressions become:

$$E[\text{event } \{N_R = 1, N_S = 1\} \text{ per patient}]$$
$$= (D_2 - y)(1 - e^{-\lambda(y-x)}) - (D_2 - x - i)\, e^{-\lambda i}\Big|_{i=y-x}^{D_2-x} - \int_{y-x}^{D_2-x} e^{-\lambda i}\, \mathrm{d}i$$
$$= D_2 - y - \frac{1}{\lambda}(e^{-\lambda(y-x)} - e^{-\lambda(D_2-x)})\ .$$

And for $y < x$:

$$E[\text{event } \{N_R = 1, N_S = 1\} \text{ per patient}] = \int_0^{D_2-x} (D_2 - i - x)\lambda e^{-\lambda i}\, \mathrm{d}i$$
$$= -(D_2 - i - x)e^{-\lambda i}\Big|_{i=0}^{D_2-x} - \int_0^{D_2-x} e^{-\lambda i}\, \mathrm{d}i$$
$$= (D_2 - x) - \frac{1}{\lambda}(1 - e^{-\lambda(D_2-x)})\ .$$

The expected cycle length is the same as in the previous section, thus we may conclude:

$$P\{N_R = 1, N_S = 1, S^{res} \geq x, R^{res} \geq y\} = \begin{cases} \dfrac{\left[D_2 - y - \frac{1}{\lambda}(e^{-\lambda(y-x)} - e^{-\lambda(D_2-x)})\right] e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda})e^{\lambda D_2}} & \text{for } y \geq x \\[2ex] \dfrac{\left[(D_2 - x) - \frac{1}{\lambda}(1 - e^{-\lambda(D_2-x)})\right] e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda})e^{\lambda D_2}} & \text{for } y < x \end{cases}$$

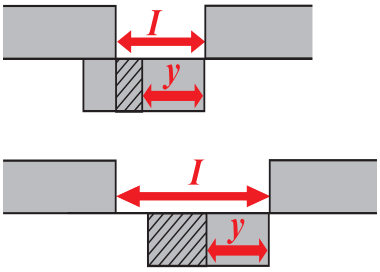


Figure 5.4: Possibilities for the length of interval $\{N_R = 1, N_S = 0, R^{res} \geq y\}$.

5.3 The event $\{N_R = 1, N_S = 0\}$

Here we obtain an expression for $P\{N_R = 1, N_S = 0, R^{res} \geq y\}$ with $0 \leq y < D_2$. Again for an arbitrary service time in the renewal cycle, we can have the following two situations (see Figure 5.4):

- $y < I \leq D_2$, then the event $\{N_R = 1, N_S = 0\}$ has length $I - y$,
- $I > D_2$, then the event $\{N_R = 1, N_S = 0\}$ has length $D_2 - y$.

Conditioning on the interarrival time gives the following:

$$
\begin{aligned}
E[\text{event } \{N_R = 1, N_S = 0\} \text{ per patient}] &= \int_y^{D_2} (i-y)\lambda e^{-\lambda i} di + \int_{D_2}^{\infty} (D_2 - y)\lambda e^{-\lambda i}\,\mathrm{di} \\
&= -(i-y)e^{-\lambda i}|_{i=y}^{D_2} - \int_y^{D_2} -e^{-\lambda i}\,\mathrm{di} + (D_2 - y)e^{-\lambda D_2} \\
&= \frac{1}{\lambda}(e^{-\lambda y} - e^{-\lambda D_2})\ .
\end{aligned}
$$

Therefore we can conclude:

$$
P\{N_R = 1, N_S = 0, R^{res} \geq y\} = \frac{\frac{1}{\lambda}(e^{-\lambda y} - e^{-\lambda D_2})e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda})e^{\lambda D_2}}
$$

5.4 The event $\{N_R = 0, N_S = 0\}$

The event $\{N_R = 0, N_S = 0\}$ occurs only when $I > D_2$. When this event occurs, the length of the interval is always $I - D_2$. Therefore we can say:

$$
\begin{aligned}
E[\{N_R = 0, N_S = 0\} \text{ per patient}] &= \int_{D_2}^{\infty} (i - D_2)\lambda e^{-\lambda i}\,\mathrm{di} \\
&= -(i - D_2)e^{-\lambda i}|_{i=D_2}^{\infty} - \int_{D_2}^{\infty} -e^{-\lambda i}\,\mathrm{di} \\
&= \frac{1}{\lambda}e^{-\lambda D_2}\ .
\end{aligned}
$$

And from this we can conclude:

$$\begin{aligned} P\{N_R = 0, N_S = 0\} &= \frac{\frac{1}{\lambda} e^{-\lambda D_2} e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}} \\ &= \frac{\frac{1}{\lambda}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}} \ . \end{aligned}$$

5.5 Summary stationary probabilities

The following stationary probabilities define the ordinary reservation model:

1. $P\{N_R = 0, N_S = 1, S^{res} \geq x\} = \frac{[D_1 - D2 - \frac{1}{\lambda}(e^{-\lambda(D_2 - x)} - 1)] e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}$ for $0 \leq x < D_1$

2. $P\{N_R = 1, N_S = 1, S^{res} \geq x, R^{res} \leq y\} =$

   $\frac{[D_2 - y - \frac{1}{\lambda}(e^{-\lambda(y-x)} - e^{-\lambda(D_2 - x)})] e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}$ for $0 \leq x \leq y < D_2$

   $\frac{[(D_2 - x) - \frac{1}{\lambda}(1 - e^{-\lambda(D_2 - x)})] e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}$ for $0 \leq y < x < D_2$

3. $P\{N_R = 1, N_S = 0, R^{res} \geq y\} = \frac{\frac{1}{\lambda}(e^{-\lambda y} - e^{-\lambda D_2}) e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}$ for $0 \leq y < D_2$

4. $P\{N_R = 0, N_S = 0\} = \frac{\frac{1}{\lambda}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}$

We can show the reservation model does not have a product-form stationary distribution by compairing $P\{N_R = 0, N_S = 0\}$ with $P\{N_R = 0\} \times P\{N_S = 0\}$. When the stationary distribution is of product-form these two expressions must be equal. Therefore we state:

Theorem 2. *The reservation model does not have a product-form stationary distribution.*

Proof. We give a direct proof of this theorem by calculating $P\{N_R = 0\} \times P\{N_S = 0\}$ and showing this is not equal to $P\{N_R = 0, N_S = 0\}$.
In the same way as in the previous sections, we can see that per patient an average of $\frac{1}{\lambda}$ of the time the event $\{N_S = 0\}$ occurs. Therefore we conclude:

$$P\{N_S = 0\} = \frac{\frac{1}{\lambda}}{D_1 + \frac{1}{\lambda}}.$$

Moreover the expected time per patient the event $\{N_R = 0\}$ occurs, is $D_1 + \frac{1}{\lambda} - D_2$. Therefore we conclude:

$$P\{N_R = 0\} = \frac{[D_1 + \frac{1}{\lambda} - D_2] e^{\lambda D_2}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}.$$

From the previous subsection we know:

$$P\{N_R = 0, N_S = 0\} = \frac{\frac{1}{\lambda}}{(D_1 + \frac{1}{\lambda}) e^{\lambda D_2}}.$$

From these three results it is not difficult to see $P\{N_R = 0, N_S = 0\} \neq P\{N_R = 0\} \times P\{N_S\}$ which proves the theorem. □

Although the reservation model does not have a product-form stationary distribution, in the next chapter we show the model has some interesting features. These features we have found by simulation and we prove some of them analytically. The simulation results and proofs are given in the remainder of this report.

Chapter 6

Simulation study

Because the impact of reservation on the blocking probability is non-trivial, we have build a simulation model to investigate this impact. In this chapter we first describe all features of the simulation model and hereafter we give some results of the simulation study.

6.1 Model description

In order to compare the behaviour of the reservation model and the $M|G|c|c$ queue, we have build a discrete event simulation model in C++. This model consists of two parts: the $M|G|c|c$ part and the reservation part. In the first part patients arrive according to a Poisson process and are accepted in case the number of patients in the queue is less than the capacity c. When an arriving patient is accepted, a service requirement is drawn from distribution G and the departure time of this patient is stored. Hereafter the next arrival moment is drawn from the exponential distribution and the next departure time is determined. When the next event is an departure, the next departure time is determined if there are still patients left in the system or set to a large number if there is nobody left. From this simulation we store the blocking probability of each run.

The second part of the simulation model consists of the reservation model. In the reservation model three events can occur: an arrival to the reservation queue, a departure from the reservation queue into the service queue and a departure from the system. When an arrival to the reservation queue occurs, an arriving patient first draws his reservation and service time from the appropriate distributions and then it is checked whether he is blocked or not. If the patient is blocked, a next arrival time is drawn from the exponential distribution. If the patient is accepted, his reservation and service requirements are stored into the state vector and the next arrival/departure time at both queues is determined. When there is a departure at one of the queues, the next departure time for that queue is either determined or set to a large number. Also for the reservation model, we store the blocking probability of each run.

In order to make a fair comparison, we applied common random numbers (see [24]) to both parts of the simulation model. In this way the results of the reservation model relative to the non-reservation model, are always reliable. Moreover we know already what the blocking probability is in the $M|G|1|1$ queue with expexted service time $\mathbb{E}(S)$ and λ the parameter of the arrival process (for example from [1]);

$$\frac{\mathbb{E}(S)}{\frac{1}{\lambda} + \mathbb{E}(S)}.$$

And in general, the blocking probability in a $M|G|c|c$ queue is given by:

$$\frac{\frac{(\lambda\mathbb{E}(S))^c}{c!}}{\sum_{n=0}^{c}\frac{(\lambda\mathbb{E}(S))^n}{n!}}.$$

Comparing the analytic blocking probability with the result of the simulation, we decided 500 simulation runs and per run 10000 patients arriving to the queue(s) resulted in 3 significant digits and a reasonable computation time. In these simulations we did not use initial-data deletion (excluding some results in the beginning of each simulation run in order to use only the results when the queue is stationary, see [24]), because the blocking probability resulting from the simulations is close to the theoretic value.

When the simulation is completed, the result is a vector of length 500 containing the blocking probability of each run for both models. We then calculate the mean blocking probability and the mean difference of the blocking probability for both models. For the difference we also compute the variance and the 90%, 95% and 99% confidence intervals.

In the next sections some results of the simulation study are described. In the next chapter we describe the results in the form of a theorem and provide proofs.

6.2 Deterministic service requirements

From the simulation study the blocking probability in the $M|D|1|1$ queue appears to be smaller than or equal to the blocking probability in the reservation model with deterministic service requirements. This result holds for all distributions of the reservation time. Moreover we see that the difference between the models increases with λ (and therefore with $\rho = \frac{\lambda}{\mu c}$ with c the capacity of the service queue).
A few examples can be seen in Figure 6.1. In this figure the blocking probability for three different reservation distributions are given together with the blocking probability of the $M|D|1|1$ queue. In all cases there was only 1 server, the service time is deterministic (with $S = 0.25$) and for the three reservation distributions $\mathbb{E}(\mathbb{R}) = 0.5$. Uniform reservation refers to R uniformly distributed on $[0, 2\mathbb{E}(\mathbb{R})]$ and for discrete reservation we took $R = 0.1$, 0.2 or 1.2, all with equal probability. We have chosen these distributions such that one coefficient of variation is smaller than 1 (the uniform distribution), one is larger than 1 (the discrete distribution) and one is equal to 1 (exponential distribution).

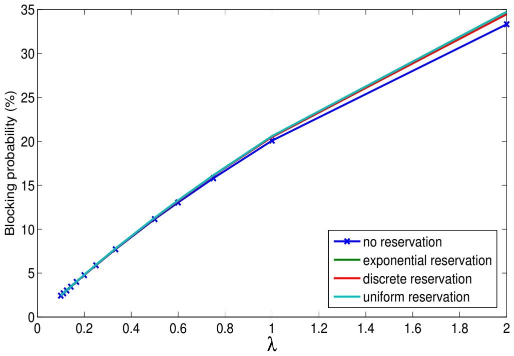


Figure 6.1: Deterministic service requirements.

In Figure 6.1 it can be seen that the blocking probability for low values of λ is almost equal for the reservation and non-reservation queue. However if we look at the 99% confidence intervals of the difference between for example the exponential reservation queue and the $M|D|1|1$ queue, then for $\lambda > \frac{1}{9}$ the blocking probability of the reservation queue is already significantly higher (0 is not in the confidence interval of the difference) than in the $M|D|1|1$ queue.

6.3 Deterministic reservation requirements

When the reservation time is deterministic, we observe that the blocking probability is the same in both models for all capacity constraints and service distributions. Since each patient in the reservation queue has deterministic reservation time, all patients arrive to the $M|G|1|1$ queue a deterministic time R later in the reservation model than they do in the non-reservation model. All further features of the models are equal and since all patients have delay R, the blocking probability in both models is the same.

6.4 Exponential reservation and service requirements

From the theory of online scheduling, letting available jobs wait before processing them seems inefficient. The first section of this chapter can confirm this and ones intuition may say that this result will hold for all types of service distributions. Remarkably, this is not true. When the coefficients of variation of the service and reservation time distribution are large enough, the blocking probabilty in the single-server reservation model is lower than the blocking probability in the non-reservation model. In Figure 6.2 this result can be seen for exponential reservation and servive times.

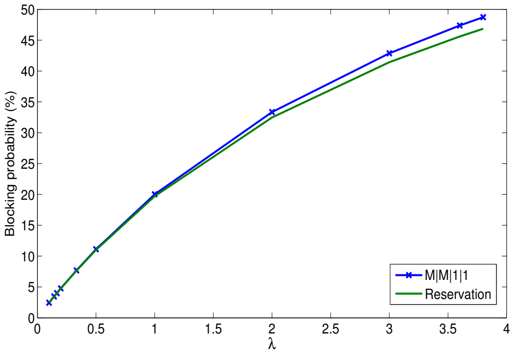


Figure 6.2: Exponential reservation and service requirements.

Also in this case it can be seen the two blocking probabilities are almost equal for low values of λ and the difference increases when λ increases. From the simulation study it appears that by letting arriving patients (with exponential service requirements) wait a random time before entering service, the mean service time of an accepted patient decreases. In the next chapter we provide an outline of the proof of this observation.

6.5 Multiple servers

So far only single server queues were simulated. In this section we present a few results of simulations with capacity 2 and 3. Because the computation time of the simulations for capacity larger than one increased rapidly, we only ran simulations for exponential reservation time and either

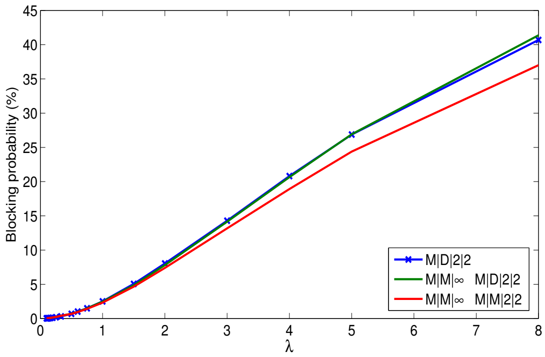


Figure 6.3: Blocking probability for capacity 2.

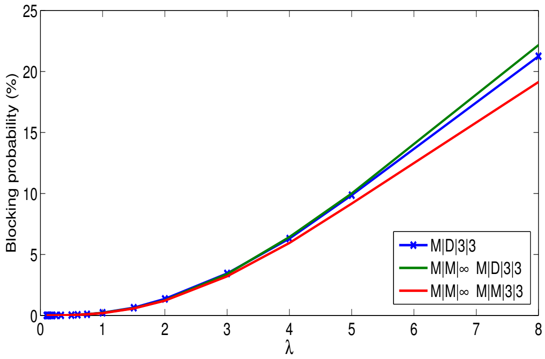


Figure 6.4: Blocking probability for capacity 3.

exponential or deterministic service time.

In Figure 6.3 it can be seen that only the case with exponential service time in the reservation model differs from the other two cases. The tag "$M|M|\infty$ $M|D|2|2$" represents the reservation model with exponential reservation time and deterministic service time. Since the loss queue is insensitive to the service distribution (see for example [34]), the blocking probability for the $M|D|2|2$ queue is equal to the blocking probability of the $M|M|2|2$ queue.
The reservation model with deterministic service requirements has a blocking probability almost equal to the $M|D|2|2$ blocking probability. For high values of λ the blocking probability is significantly higher (zero is not in the confidence interval) than the blocking probability in the $M|D|2|2$ queue, but this difference is small. For the exponential service requirement there is however a relatively large difference between the reservation and non-reservation model. This difference is about the same as in the single-server case for the same ρ with $\rho = \frac{\lambda}{\mu c}$ and c the capacity of the service queue. The previous results were also found for capacity 3, as can be seen in Figure 6.4.

In the next chapter we formulate a theorem for the results we described in the first three sections of this chapter and we provide a proof of the first two theorems. The third theorem we were unable to prove and therefore we will only provide the structure the prove should have. For the multiple server-case we only prove the results for the deterministic reservation or service requirements.

Chapter 7

Analytical derivation of the blocking probability

In the previous chapter we described several interesting cases of the reservation model. In this chapter we summarize the results of the first three sections of the previous chapter into a theorem and provide a proof of the first two theorems. The third section is on the exponential case and contains only an outline of the proof. The sections of this chapter are in the same order as in the previous chapter, with the exception that in this chapter the first two sections also include multiple server queues.

7.1 Deterministic service requirements

From the simulation study it appears that when the service requirement is deterministic, the blocking probability in the reservation model with c servers in the second queue (hereafter: c-server reservation model) is always larger than or equal to the blocking probability in the $M|D|c|c$ queue. We therefore define and prove the following theorem.

Theorem 3. *The blocking probability of the c-server reservation model with deterministic service requirements is greater than or equal to the blocking probability in the $M|D|c|c$ queue.*

Before we prove this theorem, we clarify the Greedy On-line Algorithm (GOL-algorithm) as described by Faigle and Nawijn in [15]. The GOL-algorithm is designed for scheduling intervals on-line. The authors consider an interval lost if it is not serviced immediately (at the time of its left endpoint) and uninterruptedly. Moreover in [15] it is proven that GOL is optimal for minimizing the number of losses.

The GOL algorithm schedules jobs (in our model: patients) at arrival to any free station (at most k, which is in our case the capacity of the second queue). If all stations are busy and there is a job being processed for which the residual time is larger than the service time of the arriving job, then this job is interrupted (and lost) and the new job is scheduled at that station. We clarify the GOL-algorithm by means of Figure 7.1. In this figure a realisation of an infinite server queue is given. Here each block represents one patient and the numbers on the left are the numbers

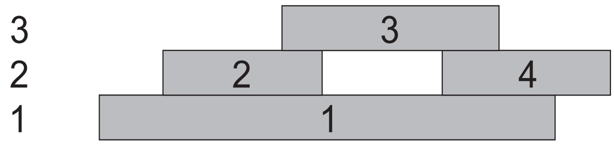


Figure 7.1: Realisation of infinite server queue.

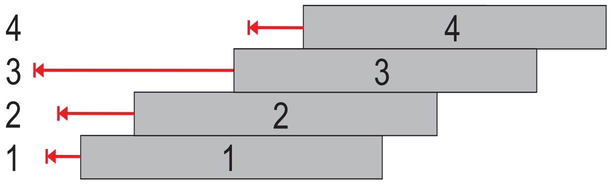


Figure 7.2: Realisation of $M|D|\infty$ queue.

of the servers that process the patients. When we apply the GOL-algorithm *for capacity 1* on the patients in Figure 7.1 we would first accept patient 1, then interrupt and block patient 1 and accept patient 2, block patient 3 and accept 4. So, we accept patients 2 and 4, and block patients 1 and 3.
With this algorithm we can prove Theorem 3.

Proof. Suppose we know exactly N patients arrive in a time window $[0, T]$, with $N \in \mathbb{N}$ and $T \in \mathbb{R}$. We define a set A^S containing the accepted patients when scheduled according to rule S, and define the blocking probability as the fraction of the N patients that is not served uninterruptedly. Consider an infinite server queue in which arriving patients are assigned a server with a number as low as possible, thus server 1 has priority of serving arriving patients. Then we can interpret server 1 independently from the other servers as a loss queue with capacity 1 and in this interpretation all patients processed by other servers are lost.

In Figure 7.2 a realisation of the $M|D|\infty$ queue is given and the reservation time of each patient is represented by a red line. When we would apply the GOL-algorithm for capacity 1 to these patients without reservation time, patient 1 would be accepted and the others would be blocked and thus $A^{GOL(1)} = \{1\}$ with $GOL(k)$ the GOL-algorithm applied for capacity k. Since the service requirements are deterministic, an arriving patient can never have a remaining service time lower than the patient already in the system.
When patients 1, 2, 3 and 4 would arrive to a $M|D|1|1$ queue, only patient 1 would be accepted and thus $A^{M|D|1|1} = \{1\}$. In general we can say that both the GOL-algorithm for c-servers and the $M|D|c|c$ queue would accept the patients that arrive when at least one server is free and all patients that arrive when all servers are occupied are blocked. From this we can conclude that $A^{GOL(c)} = A^{M|D|c|c}$. Since Faigle and Nawijn ([15]) prove that the GOL-algorithm is optimal for minimizing the number of losses, we conclude that $|A^{M|D|c|c}|$ is the maximum number of patients that could be accepted.

For the reservation model with deterministic service requirements, it depends on the reservation times of each patient which patient is accepted. As can be seen in Figure 7.2 in this example we would accept patient 3 and block the others. Therefore it could occur that $A^{res(1)} \neq A^{M|D|1|1}$ (with $res(j)$ the reservation model with capacity j), and in general $A^{res(c)} \neq A^{M|D|c|c}$. Since $A^{M|D|c|c}$ is optimal, we conclude that $|A^{res(c)}| \leq |A^{M|D|c|c}|$. As a consequence the number of blocked patients in the reservation model is larger than or equal to the number of blocked patients in the loss queue. This completes the proof. □

7.2 Deterministic reservation requirements

When the reservation time is deterministic, we observe that the blocking probability is the same in both models. We therefore state the following theorem.

Theorem 4. *The blocking probability of the c-server reservation model with deterministic reservation requirements is equal to the blocking probability of the $M|G|c|c$ queue.*

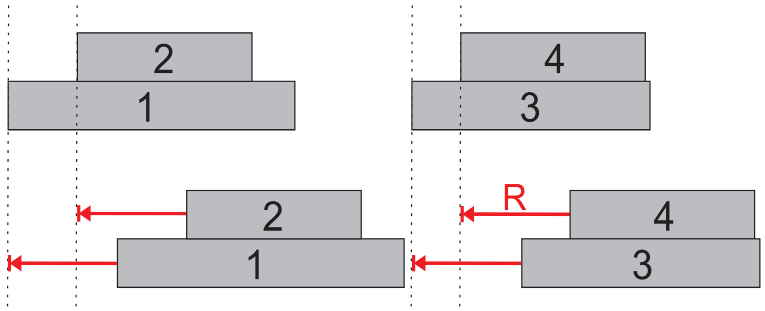


Figure 7.3: Realisation of $M|G|1|1$ queue (top) and reservation model (bottom) with blocked patients

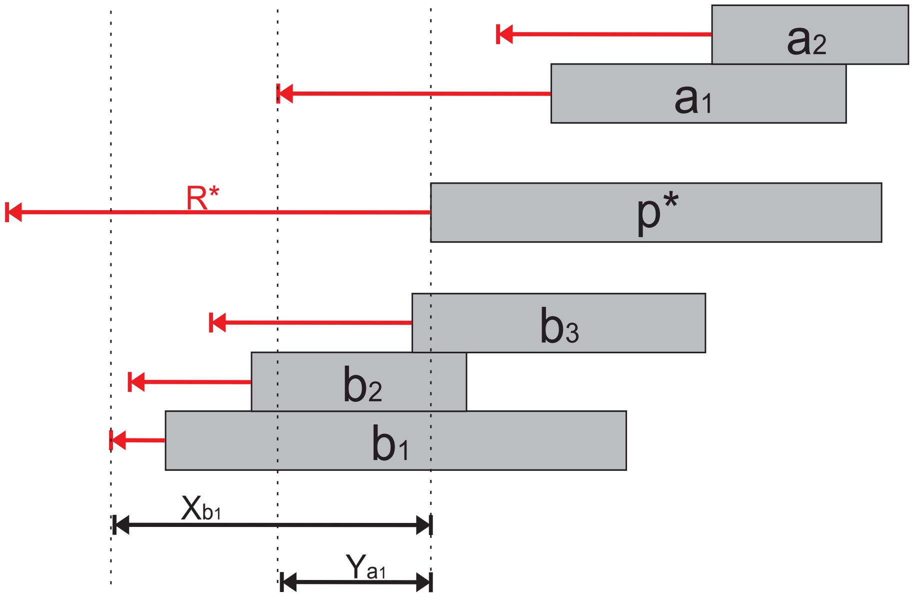


Figure 7.4: Realisation of $M|M|\infty$ queue

Proof. Proving this theorem is trivial since the c-server reservation model is equal to the $M|G|c|c$ queue where all arrivals are delayed with the reservation time R. In Figure 7.3 it can be seen that exactly the same patients are accepted (namely 1 and 3) in both models and the only difference is the delay of deterministic time R before entering service in the reservation model. □

7.3 Exponential reservation and service requirements

From the simulation study we concluded that the reservation model can work out better in terms of blocking probability. Without further introduction, we state:

Theorem 5. *The blocking probability of the 1-server reservation model with exponential reservation and service requirements is smaller than or equal to the blocking probability of the $M|M|1|1$ queue.*

In order to prove this theorem, we calculate the probability a patient is accepted in the reservation model. This probability we derive by comparing the reservation model with an infinite server queue; we calculate the probability that one specific patient "beats" the patients he finds at his arrival and the patients arriving after him in the competition for being accepted. In Figure

7.4 we see that patient p^* is accepted instead of the patients that arrived *b*efore him (b_1, b_2 and b_3) and *a*fter him (a_1 and a_2).
In order to derive the probability patient p^* is accepted, we introduce the following definitions:

- n: the number of patients present when patient p^* arrives,
- m: the number of patients arriving during the service of patient p^*,
- S_j^e: the elapsed service time, $j = b_1, b_2, ..., b_n$,
- R_j: the reservation time, $j = b_1, b_2, ..., b_n, a_1, a_2, ..., a_m$,
- I_j: the time between the arrival of patient p^* and patient j, $j = a_1, a_2, ..., a_m$,
- S^*: the service time of patient p^*,
- $X_j = S_j^e + R_j$, $j = b_1, b_2, ..., b_n$,
- $X = \max\limits_{b_1 \leq j \leq b_n} X_j$,
- $Y_j = R_j - I_j$, $j = a_1, a_2, ..., a_m$,
- $Y = \max\limits_{a_1 \leq j \leq a_m} Y_j$.

In the remainder of this section we first derive the density functions of X_j and Y_j. From this we obtain the density functions of X and Y and we derive the probability that patient p^* is accepted; $P(R^* > X, R^* > Y)$.

In this section we have assumed the reservation time R_j and service time S_j are exponentially distributed. Due to the memoryless property S_j^e is also exponentially distributed. The convolution X_j can be obtained by filling in the general formula for the density function of the convolution of random variables A and B:

$$f_{A+B}(t) = \int_{-\infty}^{\infty} f_A(z) f_B(t-z)\,\mathrm{d}z. \tag{7.1}$$

Thus, filling in this formula with $R_j \sim \exp(\mu_R)$ and $S_j^e \sim \exp(\mu_S)$ we obtain:

$$\begin{aligned} f_{X_j}(t) &= \int_0^t \mu_R e^{-\mu_R z} \mu_S e^{-\mu_S (t-z)} \,\mathrm{d}z \\ &= \mu_R \mu_S \int_0^t e^{-(\mu_R - \mu_S) z} e^{-\mu_S t} \,\mathrm{d}z \\ &= \mu_R \mu_S e^{-\mu_S t} \left[\frac{-1}{\mu_R - \mu_S} e^{-(\mu_R - \mu_S) z} \Big|_{z=0}^{t} \right] \\ &= \frac{\mu_R \mu_S e^{-\mu_S t}}{\mu_S - \mu_R} (e^{-(\mu_R - \mu_S) t} - 1) \\ &= \frac{\mu_R \mu_S}{\mu_S - \mu_R} (e^{-\mu_R t} - e^{-\mu_S t}). \end{aligned}$$

Here the boundaries of the first integral follow from the fact that the exponential distribution only exists for positive values of z and $(t-z)$. From this we conclude the density function of X_j is given by:

$$\begin{aligned} F_{X_j}(x) &= \int_0^x \frac{\mu_R \mu_S}{\mu_S - \mu_R} (e^{-\mu_R t} - e^{-\mu_S t}) \,\mathrm{d}t \\ &= \frac{-\mu_S}{\mu_S - \mu_R} e^{-\mu_R t} + \frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S t} \Big|_{t=0}^{x} \\ &= \frac{-\mu_S}{\mu_S - \mu_R} e^{-\mu_R x} + \frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \left(\frac{-\mu_S}{\mu_S - \mu_R} + \frac{\mu_R}{\mu_S - \mu_R} \right) \\ &= \frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S - \mu_R} e^{-\mu_R x} + 1. \end{aligned}$$

The density function of Y_j we can derive in a same way, only here we take the convolution of R_j with $-I_j$ and have to condition on S^*. We take $-I_j$ since the reservation time of the patients arriving after p^* needs to be larger than the time between the arrival of p^* and j, $(j = a_1, ..., a_m)$ before j has a positive probability for being accepted instead of p^*. For example, the reservation time of patient a_2 in Figure 7.4 is smaller than I_j and therefore Y_{a_2} is negative and this patient has probability zero of being accepted instead of p^*.

Since the arrival process is Poisson and we condition on m arrivals occuring during the service time S^* of p^*, we know I_j is uniformly distributed on $[0, S^*]$ (see for example Theorem 2.3.1 in [31]). Therefore we fill in (7.1) in the following way:

$$f_{Y_j|S^*}(t) = \begin{cases} \displaystyle\int_0^{S^*+t} f_{R_j}(z) f_{-I_j}(t-z) \,\mathrm{d}z & \text{for } t < 0 \\ \displaystyle\int_t^{S^*+t} f_{R_j}(z) f_{-I_j}(t-z) \,\mathrm{d}z & \text{for } t \geq 0 \end{cases}$$

For the boundaries of the integral we used that $z \geq 0$ for R_j and $-S^* \leq t - z \leq 0$ for $-I_j$. Note that $t \geq -S^*$, since this is the lower bound on Y_j.

For $-S^* \leq t < 0$, $f_{Y_j|S^*}$ becomes:

$$\begin{aligned} f_{Y_j|S^*}(t) &= \int\limits_0^{S^*+t} f_{R_j}(z) f_{-I_j}(t-z)\,\mathrm{d}z \\ &= \int\limits_0^{S^*+t} \mu_R e^{-\mu_R z} \frac{1}{S^*}\,\mathrm{d}z \\ &= \frac{-1}{S^*} e^{-\mu_R z}\Big|_{z=0}^{S^*+t} \\ &= \frac{1}{S^*}\left(1 - e^{-\mu_R(S^*+t)}\right). \end{aligned}$$

And, in the same way we obtain for $t \geq 0$:

$$\begin{aligned} f_{Y_j|S^*}(t) &= \int\limits_t^{S^*+t} f_{R_j}(z) f_{-I_j}(t-z)\,\mathrm{d}z \\ &= \int\limits_0^{S^*+t} \mu_R e^{-\mu_R z} \frac{1}{S^*}\,\mathrm{d}z \\ &= \frac{-1}{S^*} e^{-\mu_R z}\Big|_{z=t}^{S^*+t} \\ &= \frac{1}{S^*}\left(e^{-\mu_R t} - e^{-\mu_R(S^*+t)}\right). \end{aligned}$$

And therefore the conditional density function of Y_j can be derived in the following way:

$$F_{Y_j|S^*}(y) = \begin{cases} \int\limits_{-S^*}^{y} \frac{1}{S^*}\left(1 - e^{-\mu_R(S^*+t)}\right)\mathrm{d}t & \text{for } -S^* \leq y < 0 \\ \int\limits_{-S^*}^{0} \frac{1}{S^*}\left(1 - e^{-\mu_R(S^*+t)}\right)\mathrm{d}t + \int\limits_{0}^{y} \frac{1}{S^*}\left(e^{-\mu_R t} - e^{-\mu_R(S^*+t)}\right)\mathrm{d}t & \text{for } y \geq 0 \end{cases}$$

$$= \begin{cases} \frac{1}{S^*}\left(t + \frac{1}{\mu_R} e^{-\mu_R(S^*+t)}\Big|_{t=-S^*}^{y}\right) & \text{for } -S^* \leq y < 0 \\ \frac{1}{S^*}\left(t + \frac{1}{\mu_R} e^{-\mu_R(S^*+t)}\Big|_{t=-S^*}^{0}\right) + \frac{1}{S^*}\left(\frac{-1}{\mu_R}(e^{-\mu_R t} - e^{-\mu_R(S^*+t)})\Big|_{t=0}^{y}\right) & \text{for } y \geq 0 \end{cases}$$

And from this we conclude:

$$F_{Y_j|S^*}(y) = \begin{cases} \frac{1}{S^*}\left(S^* + y + \frac{1}{\mu_R}\left(e^{-\mu_R(S^*+y)} - 1\right)\right) & \text{for } -S^* \leq y < 0 \\ 1 + \frac{1}{\mu_R S^*}\left(e^{-\mu_R(S^*+y)} - e^{-\mu_R y}\right) & \text{for } y \geq 0 \end{cases}$$

Now we have obtained the expressions for the (conditional) density functions of X_j and Y_j, we derive the (conditional) density functions of X and Y. For the density function of X we use that all X_j are mutually independent and identically distributed, and thus

$$\begin{aligned} P\left(\max_{b_1 \leq j \leq b_n}(X_j) \leq x\right) &= P(X_{b_1} \leq x, X_{b_2} \leq x, ..., X_{b_n} \leq x) \\ &= P(X_{b_1} \leq x) \cdot P(X_{b_2} \leq x) \cdots P(X_{b_n} \leq x), \end{aligned}$$

and the probability that patient p^* finds n patients before him is equal to the stationary probability of having n patients in the $M|M|\infty$ queue, which Poisson with parameter λ/μ_S. From this we conclude:

$$P(X = 0) = e^{-\frac{\lambda}{\mu_S}},$$

since this is the probability that patient p^* finds the infinite server queue empty. When there are patients before p^* it holds:

$$\begin{aligned} P(X \leq x) &= \sum_{n=1}^{\infty} e^{-\frac{\lambda}{\mu_S}} \frac{\left(\frac{\lambda}{\mu_S}\right)^n}{n!} P\left(X_j \leq x\right)^n \\ &= \sum_{n=1}^{\infty} e^{-\frac{\lambda}{\mu_S}} \frac{\left(\frac{\lambda}{\mu_S}\right)^n}{n!} \left(\frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S - \mu_R} e^{-\mu_R x} + 1\right)^n \\ &= e^{-\frac{\lambda}{\mu_S}} \left(e^{\frac{\lambda}{\mu_S}\left(\frac{\mu_R}{\mu_S-\mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S-\mu_R} e^{-\mu_R x} + 1\right)} - 1\right) \\ &= e^{\frac{\lambda}{\mu_S}\left(\frac{\mu_R}{\mu_S-\mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S-\mu_R} e^{-\mu_R x}\right)} - e^{-\frac{\lambda}{\mu_S}}. \end{aligned}$$

Here the third line on the right hand side follows from the Taylor series of e^z;

$$e^z = \sum_{i=0}^{\infty} \frac{z^i}{i!}\ .$$

The random variables Y_j are also mutually independent and identically distributed and the number of patients arriving after patient p^* is also Poisson distributed, but with parameter λS^*. Therefore we can derive the density function of Y by conditioning on S^* as follows:

$$\begin{aligned} &P\left(Y \leq y | S^*\right) = \\ &= \begin{cases} \sum\limits_{m=0}^{\infty} e^{-\lambda S^*} \frac{(\lambda S^*)^m}{m!} \left(\frac{1}{S^*}\left(S^* + y + \frac{1}{\mu_R}\left(e^{-\mu_R(S^*+y)} - 1\right)\right)\right)^m & \text{for } -S^* \leq y < 0 \\ \sum\limits_{m=0}^{\infty} e^{-\lambda S^*} \frac{(\lambda S^*)^m}{m!} \left(1 + \frac{1}{\mu_R S^*}\left(e^{-\mu_R(S^*+y)} - e^{-\mu_R y}\right)\right)^m & \text{for } y \geq 0 \end{cases} \\ &= \begin{cases} e^{-\lambda S^*} e^{\lambda S^*\left(\frac{1}{S^*}\left(S^*+y+\frac{1}{\mu_R}\left(e^{-\mu_R(S^*+y)}-1\right)\right)\right)} & \text{for } -S^* \leq y < 0 \\ e^{-\lambda S^*} e^{\lambda S^*\left(1+\frac{1}{\mu_R S^*}\left(e^{-\mu_R(S^*+y)}-e^{-\mu_R y}\right)\right)} & \text{for } y \geq 0 \end{cases} \end{aligned}$$

$$= \begin{cases} e^{\lambda y + \frac{\lambda}{\mu_R}\left(e^{-\mu_R(S^*+y)} - 1\right)} & \text{for } -S^* \le y < 0 \\ e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R(S^*+y)} - e^{-\mu_R y}\right)} & \text{for } y \ge 0 \end{cases}$$

Now we have all necessities for deriving the conditional probability that patient p^* is accepted. Since for both X and Y the density function consists of two parts, we have to write out $P\left(R^* > X, R^* > Y|S^*\right)$ as follows:

$$P\left(R^* > X, R^* > Y|S^*\right) = P(X = 0) \cdot \left(P(Y \le 0|S^*)P(R^* > 0) + \int\limits_{0^+}^{\infty} P(R^* > y)\, \mathrm{d}F_{Y|S^*}(y) \right) + $$
$$+ \int\limits_{0^+}^{\infty} \int\limits_{-S^*}^{x} P(R^* > x)\, \mathrm{d}F_{Y|S^*}(y)\, \mathrm{d}F_X(x) + \int\limits_{0^+}^{\infty} \int\limits_{x}^{\infty} P(R^* > y)\, \mathrm{d}F_{Y|S^*}(y)\, \mathrm{d}F_X(x). \tag{7.2}$$

Here the first term on the right hand side represents the case $\{X = 0, R^* > Y\}$. Since R^* is exponentially distributed we know $R^* \ge 0$ and therefore the case $\{Y \le 0\}$ is mentioned separately. The second term on the right hand side represents the case $\{R^* > X, X > Y\}$ and the last term represents $\{R^* > Y, Y > X\}$. Since we handle the case $\{X = 0\}$ separately, the lower bound on x in the integrals is set to 0^+.

With all we have derived before, we can now specify all formulas required for calculating (7.2). First we state the expressions that were not mentioned before and after this we fill in (7.2). Since R^* is exponential, it holds:

$$P(R^* > x) = e^{-\mu_R x}\ .$$

From the density functions of X and Y we can derive $\mathrm{d}F_X(x)$ and $\mathrm{d}F_{Y|S^*}(y)$ as follows:

$$\mathrm{d}F_X(x) = f_X(x)\, \mathrm{d}x,$$

and for $X \neq 0$

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(e^{\frac{\lambda}{\mu_S}\left(\frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S - \mu_R} e^{-\mu_R x}\right)} - e^{-\frac{\lambda}{\mu_S}} \right)$$
$$= \frac{\lambda \mu_R}{\mu_S - \mu_R}\left(e^{-\mu_R x} - e^{-\mu_S x}\right) e^{\frac{\lambda}{\mu_S}\left(\frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S - \mu_R} e^{-\mu_R x}\right)}.$$

In the same way, for $y \ge 0$:

$$f_{Y|S^*}(y) = \frac{\mathrm{d}}{\mathrm{d}x} e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R(S^*+y)} - e^{-\mu_R y}\right)}$$
$$= \lambda\left(e^{-\mu_R y} - e^{-\mu_R(S^*+y)}\right) e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R(S^*+y)} - e^{-\mu_R y}\right)}.$$

Now we can write (7.2) as:

$$
\begin{aligned}
P\left(R^* > X, R^* > Y | S^*\right) = e^{-\frac{\lambda}{\mu_S}} \Bigg(& e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R S^*}-1\right)} \cdot 1 + \ldots \\
& \ldots \int_{0^+}^{\infty} e^{-\mu_R y} \cdot \lambda \left(e^{-\mu_R y} - e^{-\mu_R (S^*+y)}\right) e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R (S^*+y)} - e^{-\mu_R y}\right)} \mathrm{d}y \Bigg) \\
+ \int_{0^+}^{\infty} & e^{-\mu_R x} \cdot e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R (S^*+x)} - e^{-\mu_R x}\right)} \cdot \ldots \\
& \ldots \cdot \frac{\lambda \mu_R}{\mu_S - \mu_R} \left(e^{-\mu_R x} - e^{-\mu_S x}\right) e^{\frac{\lambda}{\mu_S}\left(\frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S - \mu_R} e^{-\mu_R x}\right)} \mathrm{d}x \\
+ \int_{0^+}^{\infty} \int_{x}^{\infty} & e^{-\mu_R y} \cdot \lambda \left(e^{-\mu_R y} - e^{-\mu_R (S^*+y)}\right) e^{\frac{\lambda}{\mu_R}\left(e^{-\mu_R (S^*+y)} - e^{-\mu_R y}\right)} \cdot \ldots \\
& \ldots \cdot \frac{\lambda \mu_R}{\mu_S - \mu_R} \left(e^{-\mu_R x} - e^{-\mu_S x}\right) e^{\frac{\lambda}{\mu_S}\left(\frac{\mu_R}{\mu_S - \mu_R} e^{-\mu_S x} - \frac{\mu_S}{\mu_S - \mu_R} e^{-\mu_R x}\right)} \mathrm{d}y \, \mathrm{d}x. \qquad (7.3)
\end{aligned}
$$

Here we used that $P(R^* > 0) = e^0 = 1$ and

$$
\begin{aligned}
\int_{0^+}^{\infty} \int_{-S^*}^{x} P(R^* > x) \, \mathrm{d}F_{Y|S^*}(y) \, \mathrm{d}F_X(x) &= \int_{0^+}^{\infty} P(R^* > x) \left(\int_{-S^*}^{x} \mathrm{d}F_{Y|S^*}(y) \right) \mathrm{d}F_X(x) \\
&= \int_{0^+}^{\infty} P(R^* > x) P(Y \leq x | S^*) \, \mathrm{d}F_X(x).
\end{aligned}
$$

As can be seen, the expressions in (7.3) depend on S^*. For determining $P\left(R^* > X, R^* > Y\right)$, we need to multiply (7.3) with the probability that S^* occurs and integrate over all possible values of S^*;

$$
P\left(R^* > X, R^* > Y\right) = \int_{0}^{\infty} P\left(R^* > X, R^* > Y | S^*\right) \cdot \mu_S e^{-\mu_S S^*} \, \mathrm{d}S^*,
$$

since S^* is exponentially distributed. Since this expressions is rather complicated, we were unable to calculate $P\left(R^* > X, R^* > Y\right)$ analytically and therefore we could not prove Theorem 5. If the probability could be calculated analytically we would need to verify that this probability is larger than the probability p^* is accepted in the $M|M|1|1$ queue, which is:

$$
\frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \frac{1}{\mu_S}}.
$$

This would be an interesting topic of further research.

Chapter 8

Conclusion

In this research we propose a so-called *reservation model* and examine its properties. In this model patients arrive to a tandem queue of an infinite server queue (the reservation queue) and a loss queue. The tandem network has an exceptional blocking rule; when a patient arrives to the reservation queue, the service requirement at both queues are drawn from the appropriate distributions and it is checked whether there are sufficient resources for the new patient in the second queue, if not the patient is blocked and lost.

In this final project three special cases of the single-server reservation model have been examined: deterministic reservation time, deterministic service time and exponential reservation and service time. In every case we compared the probability an arriving patient is blocked in the reservation model with the oridinary loss queue. In case of deterministic reservation time both models have equal blocking probability. With deterministic service requirements the probability a patient is blocked in the reservation model is greater than or equal to the blocking probability in the ordinary loss queue. When both the reservation time and service time are exponentially distributed, the reservation model has a blocking probability smaller than or equal to the blocking probability in the ordinary loss queue.

When we examined cases with capacity more than one, we found that the all effects of the reservation queue also occur when capacity increases.

Chapter 9

Discussion

This research originated from a question in a rehabilitation care institute from which we eventually got interested in a loss network with the possibility to claim resources in advance. In our research we decided to use a random reservation time, which could represent the patients wishes. (Sometimes patients ask for their treatment to start later, for example because they want to attend a wedding.) With random reservation times, it could occur that a patient is blocked while some small time in the future the required resources would be available. In future research more realistic assumptions must be made to prevent these situations to occur and make the model applicable for the rehabilitation centre.

As said earlier in this report, the proof of the "single server exponential reservation and service times"-case was not completed within this research. It would be interesting to complete the proof analytically and perhaps extend it to a multiple server case.

In this report we proved the stationary distribution of the reservation model does not have product form. It would be interesting to know how much the stationary distribution differs from the product-from stationary distribution of the tandem queue $D|D|\infty \rightarrow D|D|1|1$. If this difference could be estimated, we could probably derive more features of the reservation model.

Another interesting topic for further research would be the transition that occurs in the reservation model; for some service and reservation time distributions the reservation model performs worse than the ordinary loss queue and for some it performs better. Here the performance measure is the probability an arriving patient is blocked. The simulation study indicated there could be a phase-transition for both distributions, which would be a curious feature of the reservation model. One possibility for finding this transition is by analytically calculate the blocking probability for a larger class of distributions.

Bibliography

[1] I. Adan and J.A.C. Resing. Queueing theory. http://www.win.tue.nl/ iadan/queueing.pdf, 2002.

[2] E.M. Arkin and E.B. Silverberg. Scheduling jobs with fixed start and end times. *Discrete Applied Mathematics*, 18:1–8, 1987.

[3] B. Avi-Itzhak and M. Yadin. A sequence of two servers with no intermediate queue. *Management Science*, 11(5, March):553–564, 1965.

[4] S.A. Berezner and A.E. Krzesinski. Call queueing in circuit-switched networks. *Telecommunication Systems*, 6:147–160, 1996.

[5] T. Bonald. The erlang model with non-poisson call arrivals. *SIGMetrics/Performance06*, (June 26-30), 2006.

[6] P. Borowiecki and F. Göring. Greedymax-type algorithms for the maximum independent set problem. In Černá et al., editor, *SOFSEM 2011, LNCS 6543*, pages 146–156. Springer-Verlag Berlin Heidelberg, Berlin.

[7] R.J. Boucherie and N.M. van Dijk. On the arrival theorem for product form queueing networks with blocking. *Performance Evaluation*, 29:155–176, 1997.

[8] O. Boxma. $M|G|\infty$ tandem queues. *Stochastic Processes and their Applications*, 18:153–164, 1984.

[9] J.W. Bruggink, B. Lodder, and M. Kardal (Centraal Bureau voor de Statistiek). Gezonde levensverwachting neemt toe. http://www.cbs.nl/nr/exeres/9524BF75-F336-47F5-BD9A-C474ADCBAFAD.htm?RefererType=RSSItem, Retreived 27-02-2009.

[10] M.C.M. Busch (RIVM). Welke normen zijn er voor aanvaardbare wachttijden in de zorg?" in: Volksgezondheid toekomst verkenning, nationaal kompas volksgezondheid. http://www.nationaalkompas.nl/zorg/sectoroverstijgend/welke-normen-zijn-er-voor-aanvaardbare-wachttijden-in-de-zorg/, Retreived 2-6-2011.

[11] K.H. Chang and W.F. Chen. Admission control policies for two-stage tandem queues with no waiting spaces. *Computers & Operations Research*, 30:589–601, 2003.

[12] D. Cook. *Program evaluation and review technique. Applications in Education.* U.S. Government Printing Office, Washington, 1966.

[13] L. De Bleser, R. Depreitere, K. De Waele, K. Vanhaecht, J. Vlayen, and W. Sermeus. Defining pathways. *Journal of Nursing Management*, 14(7):553–563, 2006.

[14] A. de Bruin, R. Bekker, L. van Zanten, and G. Koole. Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research*, 178:23–43, 2010.

[15] U. Faigle and W.M. Nawijn. Note on scheduling intervals on-line. *Discrete Applied Mathematics*, 58:13–17, 1995.

[16] T.A. Feo, M.G.C. Resende, and S.H. Smith. A greedy randomized adaptive search procedure for maximum independent set. *Operations Research*, 42(5 (Sep. - Oct.)):860–878, 1994.

[17] F. Gavril. Maximum weight independent sets and cliques in intersection graphs of filaments. *Information Processing Letters*, 73:181–188, 2000.

[18] M.M. Halldórsson. Approximations of independent sets in graphs. In K. Jansen and J. Rolim, editors, *Approximation Algorithms for Combinatorial Optimization*, 1998.

[19] J.J. Harms and J.W. Wong. Performance modeling of a channel reservation service. *Computer Networks and ISDN Systems*, 27:1487–1497, 1995.

[20] W.J. Hopp and M.L. Spearman. *Factory Phisics: Foundations of Manufacturing Management.* Irwin/McGraw-Hill, New York, 2nd edition, 2001.

[21] J. Jun, S. Jacobson, and J. Swisher. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operation Research Society*, 50(2):109–123, 1999.

[22] J.E. Kelley. Critical-path planning and scheduling: Mathematical basis. *Operations Research*, 9(3):296–320, 1961.

[23] F. Kelly. Loss networks. *The annals of applied probability*, 1(3):319–378, 1991.

[24] A.M. Law. *Simulation Modeling and Analysis.* McGraw-Hill, New York, 2007.

[25] Y. Lu and A. Radovanovic. Asymptotic blocking probabilities in loss networks with subexponential demands. *Journal of Applied Probability*, 44:1088–1102, 2007.

[26] R. Nelson. Queueing networks. In *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*, chapter 10. Springer-Verlag, New York, 1995.

[27] S. Nicoloso, M. Sarrafzadeh, and X. Song. On the sum coloring problem on interval graphs. *Algorithmica*, 23:109–126, 1999.

[28] Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat. World population prospects. http://esa.un.org/unpd/wpp/index.htm, Retreived 2-6-2011.

[29] J. Roberts and K. Liao. Traffic models for telecommunication services with advance capacity reservation. *Computer Networks and ISDN Systems*, 10:221–229, 1985.

[30] K. Ross. *Multiservice loss models for broadband telecommunication networks.* Springer, London, 1995.

[31] S.M. Ross. *Stochastic Processes. 2nd ed.* John Wiley and Sons Inc, U.S., 1996.

[32] Katja Schimmelpfeng, Stefan Helber, and Steffen Kasper. Decision support for rehabilitation hospital scheduling. Diskussionspapiere der wirtschaftswissenschaftlichen fakultt der universitt hannover, Universitt Hannover, Wirtschaftswissenschaftliche Fakultt, 2010.

[33] W. Sermeus and K. Vanhaecht. Wat zijn klinische paden? *Acta Hospitalia*, 3, 2002.

[34] P.G. Taylor. Insensitivity in stochastic models. In R.J. Boucherie and N.M. Van Dijk, editors, *Queueing Networks*, pages 121–140. Springer, Dordrecht, 2011.

[35] M. Themistocleous, Z. Irani, R. OKeefe, and R. Paul. Erp problems and application integration issues: An empirical survey. In *34th Hawaii International Conference on System Sciences*, 2001.

[36] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, and N. Litvak. A survey of health care models that encompass multiple departments. *International journal of Health Management and Information*, 1(1):37–69, 2010.

[37] J.T Virtamo. A model of reservation systems. *IEEE transactions on communications*, 40(1):109–118, 1992.

[38] J.T. Virtamo and S. Aalto. Stochastic optimization of reservation systems. *European Journal of Operational Research*, 51:327–337, 1991.

[39] S. Zachary and I. Ziedins. Loss networks. In R.J. Boucherie and N.M. Van Dijk, editors, *Queueing Networks*, pages 701–726. Springer, Dordrecht, 2011.