Preemptive-resume priority queueing

Wieger Hinderks, 10157417
July 12, 2014

Bachelor Thesis

Supervisor: prof. dr. R. Núñez-Queija
Second grader: dr. T.M. Nieuwenhuizen

Korteweg-de Vries Institute for Mathematics
Faculty of Natural Sciences, Mathematics, and Informatics
University of Amsterdam
Abstract

The main goal of this thesis is to establish a mathematical analysis of a priority queueing model with the preemptive-resume priority discipline. First, the basics in queueing theory are recapitulated. Second the theory of queueing systems with interrupted service is studied and applied to model a basic priority queueing system: two priority levels and both priorities have exponentially distributed service times. The mean number of customers and mean sojourn time are computed for both priority levels. Then, this model is used to find the behaviour of customers dependent on the system parameters and when priority costs are introduced. A few different customer equilibrium situations are analysed.

Eventually, the dual priority level model is extended to an arbitrary $n \in \mathbb{N}$ different priority levels, and even to general service distributions. For every single priority level the mean number of customers and mean sojourn time are computed in terms of the system parameters. Finally, the $n$-level priority queueing model with general service times will be studied in heavy traffic. The Laplace-Stieltjes transform of the queueing time will be considered and limiting behaviour of all but the lowest priority classes are trivial. For the final class, however, the queueing time has an exponential distribution when multiplied by the factor $1 - \rho$, in heavy traffic.

Title: Preemptive-resume priority queueing
Author: Wieger Hinderks, 10157417, wieger.hinderks@student.uva.nl
Supervisor: prof. dr. R. Núñez-Queija
Second grader: dr. T.M. Nieuwenhuizen
Date: July 12, 2014

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 904, 1098 XH Amsterdam
http://www.science.uva.nl/math
Contents

Introduction 5

1. Introduction to queueing theory 7
   1.1. Basic concepts and Little’s Law ........................................ 8
   1.2. M/G/1 queue .......................................................... 9
   1.3. Dual service rate single server queue ............................... 12

2. Queueing processes with interrupted service 15
   2.1. Preemptive-resume interruptions .................................... 15
   2.2. Dual service rate priority queue .................................... 18
       2.2.1. Equilibrium distribution of the high priority queue ........ 19
       2.2.2. Mean values of the low priority queue in equilibrium .... 20

3. Customer equilibria in a priority queue 23
   3.1. Introduction to game theory ......................................... 23
   3.2. Wardrop equilibria in a priority queueing system ............... 26
       3.2.1. Customer equilibrium ......................................... 26
       3.2.2. System optimum ............................................... 28
   3.3. Including service costs in a priority queueing system .......... 29
       3.3.1. Setting a price to attain customer equilibrium ............. 30
       3.3.2. Customer equilibrium including service costs .............. 30
   3.4. Maximum profit in a priority queueing system .................... 30

4. Multiple priority levels 34
   4.1. Mean completion time and busy period ............................ 35
   4.2. Statistical equilibrium condition .................................. 37
   4.3. Mean performance measures ........................................ 38
   4.4. Order of service .................................................. 41
   4.5. General service times ............................................. 42
       4.5.1. Mean value analysis with identical service times .......... 43
       4.5.2. Non-identical service times ................................ 44

5. Heavy traffic approximation 48
   5.1. Mean queueing time in heavy traffic ............................... 48
   5.2. Laplace-Stieltjes transform of the queueing time ............... 49
   5.3. Distribution of the queueing time in heavy traffic .............. 50
Conclusion and review of the process 56
Bibliography 58
A. Popular summary 59
B. Final presentation 62
Introduction

When I started following the course Probability – Markov Chains, which was taught by prof. dr. Núñez-Queija at the University of Amsterdam, I was immediately interested in the subject. As, during the course, we were getting closer to the start of choosing the subject for our bachelor thesis, I started wondering whether I wanted to do my thesis on Markov chains, or perhaps queueing theory. Both dr. Núñez-Queija’s enthusiasm for teaching and the interesting mathematics itself finally made me decide that I should ask him to be the supervisor of my bachelor thesis. The rest is – well, so they say – history.

Markov chains are not just mathematically very interesting, but also have many applications in other sciences, such as physics, chemistry, biology or economics. The American scientist Elliott Montroll, for example, had a varied career involving applications of Markov chains in several areas. Amongst other things, he studied the Ising model in which he applied Markov chains and contributed to the mathematical understanding of Markov chains, [4].

In this thesis we will analyse the so-called preemptive-resume priority queue. A preemptive-resume priority queue is a queueing system consisting of two or more different queues with different priority levels – meaning that customers arrive and join a queue of a certain priority. However, there only is one service agent available, which will always serve the customers of the higher priority before those of lesser priority. The adjective preemptive means that at the moment of arrival of a customer the service of any customer with lower priority is interrupted. When those higher priority customers have left, the interrupted service may be resumed from where it left off at the beginning of the interruption – that is what the adjective resume means in the name of this priority discipline.

The first chapter is (a recapitulation of) an introduction to queueing theory, in which we will pay special attention to the M/G/1 queue – a queueing system that will be of utmost importance throughout the remainder of this thesis. We end the first chapter by detailing an example of a M/G/1 queue by explicitly finding its probability generating function, that is finding its distribution.

In the second chapter, we will focus on interrupted service, from which we can start thinking of letting service be interrupted by another queue and therefore creating a priority queueing model. After studying the theory of interrupted service, we will apply it to create a priority queueing model consisting of two priority levels which both have exponential service time distributions.

Chapter three is the odd one out, as we will leave queueing theory for a while and focus on how customers should divide themselves over the two priority levels of the model discussed in the previous chapter in order to attain certain equilibrium situations. We need some game theory in this chapter and the necessary ideas will be discussed.
The fourth chapter is on queueing theory again and we will extend the two-level
priority queueing model to a priority queueing system with arbitrarily many priority
levels. We will examine what conditions will guarantee stability. We will also extend
the model to not just allow exponentially but even generally distributed service times.

In the final chapter we will discuss the heavy traffic approximation for an \( n \)-level
preemptive-resume priority queueing system with generally distributed service times.
In heavy traffic the workload on the system is near its critical value. It is common to
investigate the effects on the distribution of the queueing time when we assume that the
workload tends to its maximum. We are inspired to look at the priority queueing model
in heavy traffic by John Kingman’s paper of 1962, \cite{3}, that stated that the distribution of
the queueing time for any general queue always converges to an exponential distribution
when scaled by the queue length.

The goal of this thesis is to become familiar with common techniques used in queueing
theory and to examine the preemptive-resume priority queueing model as precise as
possible in the given time frame.
1. Introduction to queueing theory

Queueing theory is the mathematical theory that investigates queueing processes as a stochastic system. In short, it is the study of queues, or waiting lines. Queueing theory aims to find useful information of a queueing system. Examples of such information are the mean sojourn time or the mean number of customers in the system.

A general way of studying a queueing process is by looking for a mathematical system that we actually do know very well, such as a Markov process. Sometimes it is needed to view the queue in a particular way, for example by only looking at the process at times when a customer leaves the system.

Finding the useful information is not always easy, because a queue is implicitly defined by five (usually assumed independent) properties. These properties only give secondary information on the queueing process as they describe the behaviour of customers and the service agents in the system, but not the distribution of the number of customer in the system as a whole. We define a queue using the following input, [2],

1. the arrival pattern of customers,
2. the service pattern of server agents,
3. the queue discipline,
4. the system capacity,
5. the number of service channels,

With the *arrival pattern of customers* we mean the way customers arrive to the system. Throughout this thesis, we assume customers arrive according to a Poisson process with given parameter, say $\lambda$. Poisson arrivals simulate randomness quite well, but a poisson stream is not always the logic choice for the arrival process – for example, a processing unit in a computer needs to do more work when someone is playing a game than when nobody is using the computer.

By the *service pattern* we mean how service is handled. Service can be offered to single customers or to customers in batches. The service time can be determined or have a probability distribution. In this thesis we use exponentially distributed service times but also general service times, that is, we only assume that the service time has some known cumulative distribution function.

The *queue discipline* is the order in which customers are selected, for example first come first serve (FCFS). The *system capacity* is the number of customers that are able to wait in the queue, one can imagine that the waiting room is not infinitely large and that there actually is a physical boundary on the number of customers that can enter the system. The *number of service channels* is, as the name suggests, the number of server agents that are able to help customers.
A notation has been introduced by David Kendall to avoid needing to describe all five properties in words every time [2]. A queueing process is described by a series of letters divided by slashes: \( A/B/X/Y/Z \). The symbols \( A \) and \( B \) indicate the interarrival time and the service time distribution respectively. The letters \( X \) and \( Y \) stand for the number of service channels and the system capacity respectively, and the queue discipline is given by the letter \( Z \). An example of a standard queueing system is the \( M/M/1/\infty/\text{FCFS} \) queue, which is treated in many of the introductory books on queueing theory.

1.1. Basic concepts and Little’s Law

We denote the number of customers in the system, that is in the queue and in service, at a time \( t \) by \( X(t) \). When we speak of the actual queue, we usually mean \( X(t) \). When the limiting distribution \( \pi_n = \lim_{t \to \infty} P(X(t) = n \mid X(0) = i) \) exists for all \( i \) and is independent of \( i \), we say that the queue reaches a statistical equilibrium distribution. This is also called the steady state distribution or the limiting distribution. Because in equilibrium the probabilities of the system having \( n \) customers are independent of time, we can also denote the queue in equilibrium by \( X \).

**Definition 1.1.** The mean expected number of customers \( L \) in a system that is in statistical equilibrium, is defined as

\[
L := \mathbb{E}[X] = \sum_{i=0}^{\infty} n\pi_n.
\]

As we want to handle general models, we assume that the system has \( c \geq 1 \) server agents available.

**Definition 1.2.** We define the mean number of waiting customers, or in other words those who are actually waiting in the queue, as

\[
L_q := \mathbb{E}[(X - c)\mathbf{1}_{X>c}] = \sum_{i=c+1}^{\infty} (n-c)\pi_n.
\]

Equilibrium distributions are a natural way of thinking of the system. When the process has been running for quite some time, we can approximate reality by the limiting distribution and upon our arrival we actually expect – by the so-called PASTA property – \( L \) people to be in the system and we expect \( L_q \) people to be standing in line in front of us.

From the start of the process, we can count and therefore number the customers that arrive, that means we can speak of customer \( i \) with \( i \in \mathbb{N} \) arbitrary. Upon arrival at a queue customer \( i \) has to wait some time. The total waiting time of customer \( i \) is called the sojourn time and denoted by \( S^{(i)} \). This sojourn time can be split up in two terms: the service time of customer \( i \) – denoted by \( T^{(i)} \) – which is assumed to have some known probability distribution, and the queueing time of customer \( i \) – denoted by \( T_q^{(i)} \) – which
is the time the customer actually stands in line. This means that $S^{(i)} = T_q^{(i)} + T^{(i)}$. It is common to assume that these random variables are independent and identically distributed for all customers, and therefore we can just denote $S = T_q + T$, leaving out the superscript.

**Definition 1.3.** We define the *expected queueing time* as $W_q := E[T_q]$, and the *expected sojourn time* as

$$W := E[S] = W_q + E[T].$$

To define the arrival process, we can use the time between the arrivals of customer $i$ and $i+1$, which we will denote with $T_I^{(i)}$. When we assume that this random variable is independent and identically distributed for all customers, we can just write $T_I$ as before. The effective arrival rate is the rate at which customers enter the system. When we assume that the system capacity is infinite and customers can not balk, this rate equals the arrival rate.

If customers arrive according to a Poisson process of rate $\lambda$, then the effective arrival rate equals $\lambda$ as well. We can easily see this when we remember that the distribution of the arrival process $A(t)$ – the Poisson process – conditional on $A(0)$ has a Poisson distribution, that is $(A(t) \mid A(0) = 0) \sim \text{Poisson}(\lambda t)$. Then we find for the interarrival time $T_I$ that

$$P(T_I > t) = P(A(t) - A(0) = 0) = \frac{\lambda^0}{0!} e^{-\lambda t} = e^{-\lambda t},$$

thus the interarrival time $T_I$ is exponentially distributed with mean $\lambda^{-1}$, and therefore the (effective) arrival rate equals $\lambda$.

Using the effective arrival rate we can state the following theorem, which is know as *Little’s Law*, or *Rule* or *Formula*.

**Theorem 1.4** (Little’s Law). In a system that is in equilibrium, the mean number of customers in the system equals the effective arrival rate $\lambda$ multiplied by the expected sojourn time

$$L = \lambda W,$$

and for the mean number of customers waiting in the queue and the mean queuing time this result also holds, $L_q = \lambda W_q$.

Although one intuitively expects the number of customers to be the number of customers that arrive per unit time multiplied by the waiting time, it is not an obvious rule as it is independent of the arrival pattern, the service pattern, and the queue discipline.

The last random variable we want to introduce is the *busy period duration* $T_b$ which is the time that the server is continuously busy. The expected busy period duration is called the *busy period* and we denote it by $B := E[T_b]$.

### 1.2. M/G/1 queue

In the literature ([2], [5]) the M/G/1 queue has been studied frequently. Results from these studies are discussed here in order to equip ourselves with the necessary tools we
need for studying a variety of queueing systems. In figure 1.1, a diagram of a so-called M/G/1 queue is shown.

In an M/G/1 queue the customers arrive according to a Poisson process of rate $\lambda$. As the service time has a general distribution, the easiest method of analysing this type of queue is by looking at the successive times $t_1, t_2, t_3, \ldots$ of customers leaving. So at $t_3$ the third customer leaves the system; that is the third customer completes service at $t_3$. We can then describe the discrete process $X_n := X(t_n)$ and we can see this is a Markov chain when we recognise that

$$X_{n+1} = (X_n - 1_{X_n>0}) + Y_{n+1},$$

where $Y_n$ denotes the number of arrivals during the $n$th service. At $t_{n+1}$, the first term (in brackets) is the number of customers you inherit from time $t_n$ and $Y_{n+1}$ is the number of customers that arrived during the service. Because of the Markov property of the arrival process – which is Poisson – we have that $Y_1, Y_2, \ldots$ are independent and identically distributed, and therefore $(X_n)_{n \geq 0}$ is a Markov chain.

As said before, the service time $T$ has a general probability distribution with cumulative distribution function $F_T(t)$. The Laplace-Stieltjes transform is given by

$$\hat{T}(s) := E[e^{-sT}] = \int_0^\infty e^{-st} dF_T(t),$$

(1.1)

where the Lebesgue-Stieltjes integral simplifies – when the density $f_T(t)$ exists – by replacing $dF_T(t) = f_T(t) \, dt$. We remark that $E[T] = \lim_{s \to 0^+} -\hat{T}'(s)$ and we will adopt the convenient notation of expressing the limit written at the right hand side of the equation above as $-\hat{T}'(0+)$. 

![Figure 1.1: Diagram of an M/G/1 queue with arrival rate $\lambda$ and service time $T$.](image)

**Definition 1.5 (Traffic intensity).** We define the traffic intensity as

$$\rho := \lambda E[T].$$

As the name suggests the traffic intensity is a measure for the workload of the system. For the queue to have a steady state solution we need the traffic intensity to be smaller than unity: $\rho < 1$.

Looking at $X_n$ in equilibrium we can also denote $X_n$ by $X$ with probability distribution $\pi_n$. We then define the probability generating function as

$$\phi(z) := E[z^X] = \sum_{i=0}^{\infty} \pi_i z^i.$$
There exists a relation between the probability generating function and the Laplace-Stieltjes transform – the Pollaczek-Khintchine Formula – which we state in the following theorem.

**Theorem 1.6** (Pollaczek-Khintchine Formula). The following relation holds

\[
\phi(z) = \frac{(1 - \rho)\hat{T}(\lambda(1 - z))}{\hat{T}(\lambda(1 - z))} - z,
\]

where \( \phi \) is the probability generating function of an \( M/G/1 \) queue in equilibrium.

For the proof of this result we refer to the literature, [5]. We remark that from this relation the transition probabilities \( \pi_n \) can be found, when Taylor expansion around \( z = 0 \) is possible. Besides, we find \( \pi_0 = \phi(0) = 1 - \rho \) and that \( L = \phi'(1-) \).

Now, analysing the queue using the tools we defined above, several results can be found. For the proofs of these results we also refer to the literature, [5], as including these does not contribute to the goal of this thesis, which is analysing the properties of specific queues.

**Theorem 1.7.** The mean number of customers in a steady state system equals

\[
L = E[X] = \phi'(1-) = \rho + \frac{\lambda^2 \hat{T}''(0+)}{2(1 - \rho)},
\]

and given that a customer enters the system in equilibrium, the mean queueing time is

\[
W_q = E[T_q] = \frac{(1 - \rho)\lambda \hat{T}''(0+)}{2(1 + \lambda \hat{T}'(0+))^2}.
\]

We remark that from the mean number of customers \( L \) and the mean queueing time \( W_q \) we can also find the expected sojourn time \( W \) and the mean number of waiting customers \( L_q \) by using Little’s Law or the equality \( W = W_q + E[T] \).

Also shown in literature, [2], [5], the busy period of an \( M/G/1 \) queue is found to equal

\[
B = E[T_b] = \frac{E[T]}{1 - \rho},
\]

which is a result derived from the recursive relation for the Laplace-Stieltjes transform of the busy period duration

\[
\hat{B}(s) = E[e^{-sT_b}] = \hat{T}\left(s + \lambda(1 - \hat{B}(s))\right),
\]

which is given here without proof as it can be found in the literature. From these two results it is easy to derive the mean value of \( (T_b)^2 \), and we find that

\[
B^{(2)} := E[(T_b)^2] = \frac{E[T^2]}{(1 - \rho)^3}.
\]

We will need these results on the busy period when we are studying priority queueing in chapter 2.
1.3. Dual service rate single server queue

In this section we are interested in a specific M/G/1 queueing process. The arrival process is a Poisson process of rate \( \lambda \). The service time has a hyperexponential distribution with rates \( \mu_1 \) and \( \mu_2 \). A fraction \( q_1 \in [0, 1] \) of the customers receives service at rate \( \mu_1 \) and the other fraction \( q_2 := 1 - q_1 \) receives service at rate \( \mu_2 \). Conditioning on the type of the customer in service, one can easily find the distribution of the service time \( T \) as

\[
F_T(t) = P(T \leq t) = q_1 \left(1 - e^{-\mu_1 t}\right) + q_2 \left(1 - e^{-\mu_2 t}\right),
\]

and from this we can easily find the mean service time

\[
E[T] = \frac{q_1}{\mu_1} + \frac{q_2}{\mu_2}.
\]

We then find the traffic intensity to equal

\[
\rho = q_1 \frac{\lambda}{\mu_1} + q_2 \frac{\lambda}{\mu_2} =: q_1 \rho_1 + q_2 \rho_2,
\]

where the most right equation defines \( \rho_1 \) and \( \rho_2 \). Of course, we assume \( \rho < 1 \) in order for an equilibrium to exist.

We compute the Laplace-Stieltjes transform \( \hat{T}(s) \) and its first two derivatives in order to be able to compute \( L \) and \( W_q \) using equation (1.3) and (1.4), so using equation (1.1) we find

\[
\hat{T}(s) = \int_0^{\infty} e^{-st} \left(q_1 \mu_1 e^{-\mu_1 t} + q_2 \mu_2 e^{-\mu_2 t}\right) dt = \frac{q_1 \mu_1}{\mu_1 + s} + \frac{q_2 \mu_2}{\mu_2 + s}.
\]  

(1.8)

Then, taking the second derivative gives

\[
\hat{T}'(s) = -\frac{q_1 \mu_1}{(\mu_1 + s)^2} - \frac{q_2 \mu_2}{(\mu_2 + s)^2} \quad \text{and} \quad \hat{T}''(s) = \frac{2q_1 \mu_1}{(\mu_1 + s)^3} + \frac{2q_2 \mu_2}{(\mu_2 + s)^3}.
\]

Now it is easy to find the mean number of customers in the system, and using the second derivative of the Laplace-Stieltjes transform \( \hat{T}''(0+) = 2(q_1/\mu_1^2 + q_2/\mu_2^2) \) we find

\[
L = \rho + q_1 \frac{\rho_1^2}{1 - \rho} + q_2 \frac{\rho_2^2}{1 - \rho}.
\]

When we take \( q_1 = 1 \), or equivalently \( q_2 = 0 \), our queue simplifies to a M/M/1 queue. Our result for \( L \) should also simplify to the mean number of customers in an M/M/1 queue. Note that taking \( q_1 = 1 \) corresponds to taking \( \rho = \rho_1 \) and \( \rho_2 = 0 \), so we compute

\[
L = \rho_1 + \frac{\rho_1^2}{1 - \rho_1} = \frac{\rho_1}{1 - \rho_1},
\]

which corresponds to the result found in the literature [2].
We are interested in the queueing time as well, and we use the derivates found for the Laplace-Stieltjes transform, \( T'(0^+) = -\left(\frac{q_1}{\mu_1} + \frac{q_2}{\mu_2}\right) \), combined with equation (1.4) to obtain
\[
W_q = \frac{\frac{1}{\mu_1}q_1 + \frac{1}{\mu_2}q_2}{1 - \rho} = \frac{q_1\rho_1 + q_2\rho_2}{\lambda(1 - \rho)}.
\]
Again plugging in \( q_1 = 0 \) to verify our result with the M/M/1 queue we find
\[
W_q = \frac{\rho_1}{\mu_1 - \lambda},
\]
which also corresponds with the value found in the literature, [2].

**Probability generating function**

We use our results from chapter 1.2, in particular equation (1.2) to find the generating function \( \phi(z) \) in terms of the Laplace-Stieltjes transform – which we computed in the previous section. From this we shall expand \( \phi \) around \( z = 0 \) in order to find the equilibrium distribution \( \pi_n \). First, we will need a compact form of \( \hat{T}(\lambda(1 - z)) \) in order to manipulate the formulas easily, thus using equation (1.8) we find
\[
\hat{T}(\lambda(1 - z)) = \frac{q_1}{1 + \rho_1(1 - z)} + \frac{q_2}{1 + \rho_2(1 - z)} = \frac{1 + \hat{\rho} - \hat{\rho}z}{(1 + \rho_1(1 - z))(1 + \rho_2(1 - z))},
\]
where \( \hat{\rho} := q_2\rho_1 + q_1\rho_2 \).

Plugging this result into equation (1.2) gives
\[
\phi(z) = (1 - \rho)\frac{(1 - z)\hat{T}(\lambda(1 - z))}{\hat{T}(\lambda(1 - z)) - z} = (1 - \rho)\frac{(1 - z)(1 + \hat{\rho} - \hat{\rho}z)}{1 + \hat{\rho} - \hat{\rho}z - z[1 + \rho_1(1 - z)][1 + \rho_2(1 - z)]}.
\]
As \( \hat{T}(0) = 1 \) we recognize that \( \hat{T}(\lambda(1 - z)) - z \) has a zero in \( z = 1 \). So continuing with the denominator only, it factorizes
\[
1 + \hat{\rho} - \hat{\rho}z - z[1 + \rho_1(1 - z)][1 + \rho_2(1 - z)] = (1 - z)\left[1 + \hat{\rho} - (\rho_1 + \rho_1\rho_2 + \rho_2)z + \rho_1\rho_2z^2\right],
\]
and we conclude that
\[
\phi(z) = (1 - \rho)\frac{1 + \hat{\rho} - \hat{\rho}z}{1 + \hat{\rho} - (\rho_1 + \rho_1\rho_2 + \rho_2)z + \rho_1\rho_2z^2}.
\]
Note that, as was expected, \( \phi(0) = \pi_0 \) and \( \phi(1) = 1 \).

We proceed by remarking that the denominator is a quadratic polynomial which can be written as the product \( \rho_1\rho_2(z - z_+)(z - z_-) \), where \( z_+ \) and \( z_- \) are the zeros of the polynomial. Computing these zeros with the quadratic formula is easy
\[
z_{\pm} = \frac{1}{2} \left[1 + \frac{1}{\rho_1} + \frac{1}{\rho_2} \pm \sqrt{\left(1 + \frac{1}{\rho_1} + \frac{1}{\rho_2}\right)^2 - 4\frac{1 + \rho}{\rho_1\rho_2}}\right],
\]
13
We now conclude that
\[ \phi(z) = \frac{1 - \rho}{\rho_1 \rho_2} \frac{1 + \hat{\rho} - \hat{\rho} z}{(z - z_+)(z - z_-)}, \]
which can be expanded as
\[ \phi(z) = 1 - \frac{\rho}{\rho_1 \rho_2} \left[ \frac{\alpha_+}{z_+ - z} + \frac{\alpha_-}{z_- - z} \right], \]
where \( \alpha_\pm \) is defined as
\[ \alpha_\pm = \frac{\hat{\rho} z_\pm - (1 + \hat{\rho})}{z_\pm - z}. \]
Finally, this leads to the Taylor expansion around 0 by using the geometric series
\[ \phi(z) = \sum_{i=0}^{\infty} \frac{1 - \rho}{\rho_1 \rho_2} \left[ \alpha_+ z_+^{-(i+1)} + \alpha_- z_-^{-(i+1)} \right] z^i, \]
or equivalently, this yields
\[ \pi_n = \frac{1 - \rho}{\rho_1 \rho_2} \left[ \alpha_+ z_+^{-(n+1)} + \alpha_- z_-^{-(n+1)} \right]. \]
Checking this relation for \( n = 0 \) indeed yields \( \pi_0 = 1 - \rho \).
We have applied our abstract understanding of M/G/1 queues to a system where customers are assigned to a service rate upon being served. In order to make this queueing system more interesting, we want to create two different waiting lines – both with their own service rate – such that customers are assigned to a service rate upon joining the system instead upon going into service. One for the service rate \( \mu_1 \) and another for \( \mu_2 \). If, in addition, the service of rate \( \mu_1 \) has priority by stating that service of type one customers interrupts service of type two, we get a whole different queueing system: a system with interrupted service. This will be the objective of the next chapter.
2. Queueing processes with interrupted service

It is not unimaginable for customers in a queueing system to experience interrupted service. One can think of a machine breaking down, or customers of higher priority entering the system. We will now take a look at results in the literature to extend our knowledge of continuous service queues to interrupted service. We will speak of regular customers as customers, and we will address the interruptions of the service of these customers as interruptions. Obviously, interruptions can be (high priority) customers that intervene with the service of the regular customers, but for generality we speak of interruptions.

We can divide interrupted service queues into three categories: postponable, preemptive-resume, and preemptive-repeat interruptions, of which we will only deal with the preemptive-resume interruptions mathematically. As the name suggests, postponable interruptions are postponed to the end of service whenever a customer is in service – so this kind of interruption does not intervene with the current service. Preemptive interruptions, however, are interruptions that do intervene with the current service. After a preemptive-resume interruption the service of the customer continues where it was left at the time of interruption. On the other hand, service has to start over again after a preemptive-repeat interruption.

2.1. Preemptive-resume interruptions

We assume that customers arrive and interruptions occur according to a Poisson stream of rate $\lambda$ and $\nu$ respectively. We are interested in the queueing process of the customers and write as before $X(t)$ for the number of regular customers in the system at time $t$. We now take $T$ to be the service time when no interruptions occur, thus $T$ is the actual service time. The service time and the time needed to complete a service are not always equal as an interruption can create additional delay.

**Definition 2.1.** We define the completion time $C$ as the time needed to complete a service with possible intermediate interruptions, thus

$$C = T + \sum_{n=0}^{N} T_{b,n},$$

where $N$ is the number of interruptions during a customer’s service and $T_{b,n}$ is the duration of the $n^{th}$ interruption (with $T_{b,0} = 0$).
As service of customers only depends on whether or not it is interrupted, it only depends on the total continuous duration of the interruptions. Later we will assume that interruptions are caused by arrival of customers of higher priority and the random variable $T_{b,n}$ corresponds to the busy period duration of the $n^{th}$ higher priority customer. As for every interruption we assume the different $T_{b,n}$ to be identically and independently distributed, we can denote the duration of an interruption with $T_{b,i}$ throughout the rest of this chapter (where the $i$ stands for interruption), while using $T_{b,c}$ to denote the busy period duration of the customers (where now the $c$ stands for customer). One can easily see this notation is convenient later in this thesis, when we will speak of both the busy period of the interruptions (which are higher priority customers then) and the busy period of the regular customers.

We define the Laplace-Stieltjes transform of the completion time $C$ as
$$
\hat{C}(s) := E[e^{-sC}].
$$
We state a relation with the Laplace-Stieltjes transform of the service time $T$, or $\hat{T}$, which is well known in the literature [1]. The result is given here without proof.

**Theorem 2.2.** The Laplace-Stieltjes transforms of $C$ and $T$ are related by
$$
\hat{C}(s) = \hat{T}(s + \nu - \nu \hat{B}_i(s)),
$$
where $\hat{B}_i(s)$ is the Laplace-Stieltjes transform of the duration $T_{b,i}$ of the interruptions.

From this relation the mean completion time can be determined by computing $-\hat{C}'(0+)$.

It is found that
$$
E[C] = E[T] (1 + \nu E[T_{b,i}]), \quad (2.1)
$$
and by taking derivatives once more that
$$
E[C^2] = E[T^2] (1 + \nu E[T_{b,i}])^2 + \nu E[T] E[T_{b,i}^2]. \quad (2.2)
$$
The following derivation gives more insight into the validity of the formula above. Observe that the expected completion time is the expected service time added to the product $\nu E[T] E[T_{b,i}]$. As $\nu$ is the average number of occuring interruptions per unit time, we see that during one service of a customer we expect $\nu E[T]$ interruptions. Each of these interruptions keeps the server occupied for an expected duration equal to the mean value: $E[T_{b,i}]$, so the product $\nu E[T] E[T_{b,i}]$ is the expected time that the service of a customer is interrupted. As we assumed that the server can resume the service of customers after interruption, we conclude that the mean time it takes for a customer to complete service – the mean completion time – is the mean service time plus the mean time the service is interrupted: $E[T] + \nu E[T] E[T_{b,i}]$. In retrospect, we could also have derived the formula for the mean completion time from a mean value analysis, as we just did.

Analogously to the M/G/1 queue we can define the traffic intensity of a queue with interruptions to be
$$
\rho := \lambda E[C] = \lambda E[T] (1 + \nu E[T_{b,i}]), \quad (2.3)
$$
which seems quite natural as it is only an increase of the uninterrupted traffic intensity by a factor $1 + \nu E[T_{b,i}]$. Remark that whenever we set the number of arriving interruptions
to zero – meaning $\nu = 0$ – that the mean completion time coincides with the mean service time and therefore this ‘new’ definition of the traffic intensity coincides with the ‘old’ one given in section 1.2. They are actually the same when we regard this queue with interruptions as an $M/G/1$ queue with the completion time $C$ as the service time.

As $\lambda$ is the mean number of customers arriving per unit time, we see that $\lambda^{-1}$ is the mean interarrival time. That means the traffic intensity is the ratio of the mean completion time over the mean interarrival time. Of course, a queue with interrupted service can only have a statistical equilibrium whenever the former is strictly smaller than the latter, which is equivalent to the statement $\rho < 1$. Whenever this condition holds, the queue $X(t)$ has an equilibrium distribution, which we denote with $X$. We will explicitly show for a priority queueing system that this condition is indeed enough, for which we refer to section 2.2.2.

There are two things we omitted in the previous discussion on the relation between the traffic intensity and the existence of the limiting distribution, which are server occupation by an interruption upon arrival of a customer and the probability of an on-going interruption. One could easily imagine a machine breaking down, while being idle, or an employee getting ill with no customers in the store – those are examples of an interruption happening in an idle period, and those could influence the existence of a limiting distribution, but the traffic intensity is independent of those, it seems.

However, we should realise that all interruptions that occur while there are customers present in the system are adding to the completion time $C$ and therefore influence the traffic intensity. The limiting distribution only exists when in general the server can process customers more quickly than the rate at which customers arrive. Therefore, the specific situation that there are interruptions present upon arrival of a customer does not contribute to the existence of the limiting distribution as the time the customer has to wait before going into service is finite when we assume that $\text{Var}(T_{b,i}) < \infty$, which we will assume from now on. After that finite waiting time interruptions are taken into account by the completion time.

The next remark we make, is that we implicitly assume that $\mathbb{P}(T_{b,i} < \infty) = 1$ in order for this argumentation to hold, as a positive probability on an infinite busy period of the interruptions keeps the server infinitely busy and customers can not be served anymore.

In the literature results for the busy period of the queue in equilibrium are found by expressing the Laplace-Stieltjes transform of the busy period of the whole system $T_b$ in the other random variables. The relation is given here without proof, as including the proof does not contribute to the main goal of this thesis and can be found in the literature [1].

**Theorem 2.3.** In a system with interrupted service that is in equilibrium, the following relation between the busy period of the server (thus the busy period by both customers and interruptions) and the busy period of merely the interruptions holds

$$B := \mathbb{E}[T_b] = \frac{\rho + \nu\mathbb{E}[T_{b,i}]}{(\lambda + \nu)(1 - \rho)},$$

where $\rho$ is the traffic intensity as defined in equation (2.3).
We now state results for the probability generating function,

\[ \phi(z) = \mathbb{E}[z^X], \]

of a queue with interrupted service that is in equilibrium. Without proof, we state a result from the literature [1].

**Theorem 2.4.** In a system that is in equilibrium with interrupted service, we find that the following relation holds

\[ \phi(z) = \phi^{(0)}(z) \frac{1}{1 + \nu \mathbb{E}[T_{b,i}]} \left( 1 + \frac{\nu}{\lambda} \frac{1 - \hat{B}_i(\lambda(1 - z))}{1 - z} \right), \]

where \( \hat{B}_i(w) \) is the Laplace-Stieltjes transform of the busy period duration \( T_{b,i} \) of the interruptions, and where we have defined

\[ \phi^{(0)}(z) := \frac{(1 - \rho)(1 - z)\hat{C}(\lambda(1 - z))}{\hat{C}(\lambda(1 - z)) - z}. \]

Remark that \( \phi^{(0)}(z) \) equals the probability generating function of an M/G/1 queue with service time \( C \).

From this relation the mean number of customers in the system can be found to equal

\[ L = \rho + \frac{\rho^2}{2(1 - \rho)} \frac{\mathbb{E}[C^2]}{(\mathbb{E}[C])^2} + \frac{\nu \lambda \mathbb{E}[T_{b,i}^2]}{2(1 + \nu \mathbb{E}[T_{b,i}])}, \]

(2.4)

which can be used to find \( W, W_q, \) and \( L_q \) as discussed in section 1.1.

### 2.2. Dual service rate priority queue

We modify the queueing system of chapter 1.3 into a system which has more possibilities with respect to adding game theoretical elements. The system is very similar to the first queueing process, but now the system has two separate queues – which means that the customers have to choose their type before entering the system.

The first queue \( X_1(t) \) is a regular M/M/1 queue with arrival rate \( \lambda_1 := q_1 \lambda \) and mean service time \( \mu_1 \). In the second queue \( X_2(t) \) customers arrive according to a Poisson process with rate \( \lambda_2 := q_2 \lambda \) and the service time has an exponential distribution with mean \( \mu_2 \) whenever there are no customers in \( X_1(t) \) during the service. However, when a customer of type one enters the system, the service of a type 2 customer is interrupted, making \( X_2(t) \) a queueing process with interrupted service depending on the busy period of \( X_1(t) \). Service of customers of type 2 is interrupted immediately and will resume from the point where it left off – as the exponential distribution is memoryless – making the interruptions of \( X_1(t) \) of the type preemptive-resume.

In short, there is only one server which gives priority to customers of type 1 and only serves customers of type 2 when there are no customers of type 1 left in the system.
Figure 2.1.: Diagram of a priority queue. Customers of type $i$ are served at rate $\mu_i$.

The server interrupts service of a customer of type 2, when a customer of type 1 enters. Figure 2.1 shows a diagram of this priority queueing system.

Now we can use the theory we already discussed to analyse this queue. First we will look at $X_1$, which is rather easy, and then we extend this examination to $X_2$. We are particularly interested in the statistical equilibrium of both queues.

### 2.2.1. Equilibrium distribution of the high priority queue

We denote the traffic intensity of type 1 customers by $\rho_1 := \frac{\lambda_1}{\mu_1}$ and we immediately see that $X_1(t)$ has an equilibrium distribution whenever $\rho_1 < 1$ as is proven in the literature [2], [5]. So we assume $\rho_1 < 1$ and continue. For the time independent system in equilibrium $X_1$, the probability distribution $\pi_n$ is known in terms of $\rho_1$ and from them many properties of $X_1$ can be found. From the literature it can be found that

$$P(X_1 = n) = \pi_n = (1 - \rho_1)\rho_1^n,$$

whenever $\rho_1 < 1$, or equivalently $\mu_1 > \lambda_1 = q_1\lambda$. From the transition probabilities we find the probability generating function as

$$\phi_1(z) = \mathbb{E}[z^{X_1}] = (1 - \rho_1) \sum_{i=0}^{\infty} (\rho_1 z)^i = \frac{1 - \rho_1}{1 - \rho_1 z} \quad (\text{for } 0 < z < 1).$$

With this result we can find many properties of the M/M/1 queue $X_1$ in statistical equilibrium, such as the expected sojourn time or the mean number of customers in the system.

Figure 2.2.: Diagram of $X_1$, which is a regular M/M/1 queue as $X_1$ does not see customers of type 2.

We will need the busy period of $X_1$ in order to be able to investigate $X_2$ in more detail, so using relations (1.5) and (1.7) found in chapter 1.2. The busy period of an
M/M/1 queue is given by
\[ B_1 = E[T_{b,1}] = \frac{1}{\mu_1 - \lambda_1}, \tag{2.5} \]
and the second moment by
\[ B_1^{(2)} = E[T_{b,1}^2] = \frac{2\mu_1}{(\mu_1 - \lambda_1)^3}. \]

Finally, we will find the Laplace-Stieltjes transform of the busy period duration \( T_{b,1} \) using the recursive relation of equation (1.6) and the fact that for a M/M/1 queue we have \( \hat{T}_1(w) = \frac{\mu_1}{w + \mu_1} \); so
\[ \hat{B}_1(w) = \frac{1}{2\lambda_1} \left( w + \lambda_1 + \mu_1 - \sqrt{(w + \lambda_1 + \mu_1)^2 - 4\lambda_1\mu_1} \right), \]
and with these results we can now focus on the equilibrium distribution of \( X_2 \), the low priority queue.

### 2.2.2. Mean values of the low priority queue in equilibrium

Instead of speaking of the service time \( T_2 \), it makes more sense to talk about the completion time \( C \), like we said before. The mean completion time is found by using equation (2.1) and the busy period of \( X_1 \) as computed in the previous section
\[ E[C] = \frac{1}{\mu_2} \left( 1 + \frac{\lambda_1}{\mu_1 - \lambda_1} \right) = \frac{1}{\mu_2} \frac{1}{1 - \rho_1}. \tag{2.6} \]

As discussed in previous chapters, we have an increased traffic intensity \( \rho_2 \) as it scales with the completion time instead of the service time, which is smaller. We define \( \rho_2' := \frac{\lambda_2}{\mu_2} \) as the uninterrupted traffic intensity of \( X_2(t) \), and with equation (2.3) we compute the adjusted traffic intensity
\[ \rho_2 = \lambda_2 E[C] = \frac{\rho_2'}{1 - \rho_1}. \]

For a limiting distribution to exist we need that \( \rho_2 < 1 \) as discussed earlier in section 2.1. However, we would explicitly show here that this was indeed enough. In section 1.3 we discussed a system where two M/M/1 queues were joined with probabilities \( q_1 \) and \( q_2 \) and given service in order of joining. That system is essentially the same as the priority system we are discussing now, only without the priority property. In section 1.3 we found that in order for equilibrium to exist, we needed the first of the following equivalent conditions
\[ \rho_1 + \rho_2' < 1 \iff \rho_2' < 1 - \rho_1 \quad (\rho_1 < 1) \iff \frac{\rho_2'}{1 - \rho_1} < 1, \]
where the last is exactly \( \rho_2 < 1 \) and the second equivalence is due to the fact that \( \rho_1 < 1 \). This equivalence even holds for all priority queueing systems where both queues are M/G/1, see section 4.2 and the discussion in section 4.5.
Throughout the remainder of this chapter we will assume \( \rho_2 < 1 \), thus that there exists a statistical equilibrium distribution. We will denote \( X_2(t) \) in equilibrium with \( X_2 \) as usual and we use theorem 2.3 to find the busy period of the low priority queue in equilibrium

\[
B_2 = \frac{\rho_2 + \lambda_1 B_1}{\lambda(1 - \rho_2)} = \frac{\rho_1 + \rho'_2}{\lambda(1 - \rho_1)(1 - \rho_2)},
\]

where we have used that \( \lambda_1 + \lambda_2 = \lambda(q_1 + q_2) = \lambda \).

We would like to compute the mean number of customers \( L_2 \) in the system, but as equation (2.4) is quite complicated we split the work. First, we compute

\[
E[C^2] = \frac{2}{\mu_2^2} \frac{1}{(1 - \rho_1)^2} + \frac{2}{\mu_1 \mu_2} \frac{\rho_1}{(1 - \rho_1)^3},
\]

so with equation (2.6) we find that

\[
\frac{E[C^2]}{(E[C])^2} = 2 + \frac{2\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1}.
\]

Now we focus on the third term of the mean number of customers in the system and find that

\[
\frac{\lambda_1 \lambda_2 E[T_{b,1}^2]}{2(1 + \lambda_1 B_1)} = \frac{\lambda_2}{\mu_1} \frac{\rho_1}{(1 - \rho_1)^2}.
\]

Putting this all together we conclude that the mean number of customers equals

\[
L_2 = \frac{\rho_2}{1 - \rho_2} + \frac{\mu_2}{\mu_1} \frac{\rho_1 \rho_2^2}{(1 - \rho_1) (1 - \rho_2)} + \frac{\lambda_2}{\mu_1} \frac{\rho_1}{(1 - \rho_1)^2}
\]

\[
= \frac{\rho_2}{1 - \rho_2} \left( 1 + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right), \tag{2.7}
\]

where we used that \( \rho_2 = \lambda_2 / (\mu_2 (1 - \rho_1)) \).

Now we will compute \( W_2, W_{2,q}, \) and \( L_{2,q} \) as we will use these results in the next chapter, where we add a game theoretical element to this queue. Using Little’s Law – theorem 1.4 – we find that the expected sojourn time is

\[
W_2 = \frac{L_2}{\lambda_2} = \frac{1}{\mu_2 (1 - \rho_1) (1 - \rho_2)} \left( 1 + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right). \tag{2.8}
\]

We define

\[
\rho := \rho_1 + \rho'_2 = \lambda \left( \frac{q_1}{\mu_1} + \frac{q_2}{\mu_2} \right),
\]

which is actually the same as the traffic intensity of the dual rate server system we studied in chapter 1.3. We also remark that we frequently use the following relations

\[
\frac{\rho_2}{1 - \rho_2} = \frac{\rho'_2}{1 - \rho}.
\]
and

$$(1 - \rho_1)(1 - \rho_2) = 1 - \rho$$

in manipulating the formulas. Using the definition of $\rho$ we can write

$$W_2 = \frac{1}{\mu_2(1 - \rho)} \left( 1 + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right).$$

Using the relation $W_2 = W_{2,q} + \mathbb{E}[C]$, we can deduce that the mean queueing time is

$$W_{2,q} = W_2 - \mathbb{E}[C] = \frac{1}{\mu_2(1 - \rho_1)} \left( \frac{\rho_2}{1 - \rho_2} + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right),$$

or equivalently

$$W_{2,q} = \frac{1}{1 - \rho_1} \left( \frac{\rho_2}{\mu_2(1 - \rho_2)} + \frac{\rho_1}{\mu_1(1 - \rho_1)} \right).$$

Finally, we use Little’s Law again to obtain the mean number of customers that are actually waiting in the queue

$$L_{2,q} = \lambda_2 W_{2,q} = \rho_2 \left( \frac{\rho_2}{1 - \rho_2} + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right),$$

which gives us the last of the four quantities we wanted to know in order to be able to bring in a game theoretical element.

Now that we have derived several results for the dual rate priority queueing system, we want to add some game elements to the system. Say, we ask a fee for the high priority queue and every customer is allowed to choose whether to pay this fee or not. What fraction of the customers would then go into high priority service? Or for what price would customers be indifferent whether they join the high or low priority service queue?
3. Customer equilibria in a priority queue

We are interested in the behaviour of customers, thus given specific parameters $\lambda$, $\mu_1$, and $\mu_2$ joining which queue is beneficial. What should the fractions $q_1$ and $q_2$ be in order to let customers be indifferent between which queue to join? Before looking into a few equilibrium situations, we will study some game theory.

3.1. Introduction to game theory

Game theory is the study of interaction between decision makers [6]. The theory does not only apply to games in the usual meaning – for example board games or games of sport – but does also apply to complex economic, sociologic or biologic systems in which there are several decision makers interacting with one another. The goal of game theory is understanding how these interactions influence each other, for example whether there are equilibria where everyone profits maximally. All these kind of systems of interacting decision makers are called games and the decision makers in the games are called players accordingly.

Every player has a variety of available actions to perform, and which action he or she chooses influences not only his or her position in the game, but has impact on the positions of other players as well – otherwise the game would not be very interesting as it is the interaction which creates unexpected situations. Of course it is possible for players to like some actions more than others: the preferences of players are specified at the start of the game and this can be done in two ways. The first is to specify for every player and all pairs of actions which one is preferred – doing so requires taking into account that this has to be consistent by which we mean that preference is transitive: if action $a$ is preferred to $b$ and $b$ is preferred to $c$, then $a$ is preferred to $c$ as well. The second way is to use a payoff function, which is a function, say $v$, that assigns a number to every action $a$. We then say that $a$ is preferred to $b$ if and only if $v(a) > v(b)$. Looking at these two ways of speaking of preference we see that they are actually the same and that preference is only ordinal.

**Definition 3.1.** A finite player strategic game with ordinal preferences consists of

- a set of players $\mathcal{P} = \{1, \ldots, n\}$,
- for each player $i \in \mathcal{P}$ a set of available actions $\mathcal{A}_i$,
- and for each player an (ordinal) preference relation on the set of actions.
The preference relation on $A_i$ can be represented by the ordering on $R$ with a payoff function, which for each player $i \in P$ is a linear, real-valued function $F_i : A_i \to R$.

Game theory assumes that all actions taken by players are the best choices, or at least as good as any other available action, according to their preferences. This is called the theory of rational choice [6]. It is a general principle in game theory and it is necessary to make the theory consistent, as we cannot do any quantitative analyses when players sometimes choose certain actions while there are other possible actions with higher payoff.

Strategies and game theoretical equilibria

In every situation a player, say $i \in P$, can choose to perform a single action $a \in A_i$, that is, the player adopts the pure strategy $s_i = a \in A_i$. On the other hand, the player can also choose between $N$ different actions $a_k$ with probabilities $p_k$, then the player adopts the mixed strategy $s_i = \sum_{k=1}^{N} p_k a_k$ (where $\sum_{k=1}^{N} p_k = 1$). We denote the set of all possible strategies for player $i$ with $S_i$.

Given a game with $n$ players, a strategy profile $s = (s_1, s_2, \ldots, s_n)$ is a tuple of strategies $s_i \in S_i$ with $i \in P$, such that each player $i$ chooses strategy $s_i$.

As we assume that the actions of all players influence one player, we should consider this in the definition of the payoff function. For $i \in P$ denote $s_i = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ for a strategy profile without the strategy of player $i$. We define the payoff function of player $i$ against strategy $s_i$ as a linear, real-valued function $F_i : A_i \to R$. This is the payoff function of player $i$ when the other players use the strategy profile $s_i$.

It is now natural to speak of the best response $s_i' \in S_i$ for player $i$ against the strategy profile $s_i'$ if $s_i' \in \arg \max_{s_i \in S_i} F_i(s_i, s_i')$. It feels quite intuitive to say that the players choose an equilibrium strategy profile whenever it is not profitable for any single player to deviate from this strategy profile. In regard of the theory of rational choice, nobody will change his or her strategy then.

**Definition 3.2** (Nash equilibrium). Given a game with $n$ players, we say a strategy profile $s_e = (s'_1, \ldots, s'_n)$ is a Nash equilibrium if it is the best response against itself or, in other words, for all $i \in P$ we have

$$s'_i \in \arg \max_{s_i \in S_i} F(s_i, s'_i).$$

Of course, we have only been dealing with finite player games, whereas we want to apply game theory to a queuing process with an unknown number (possibly infinitely many) of players at any given time. That is why we also need to define notions of equilibria in case we deal with an unknown number of players, when the Nash equilibrium cannot be determined.

To do so, we assume all players have the same set of available strategies $S$. This is a legitimate assumption in queueing theory, all customers or players are usually given the same choices, for example whether to join a queue or to balk. We also assume that every player has the same preferences, because in statistical equilibrium we cannot distinguish
between customers. We write $F(s, s_0)$ for the payoff function of an arbitrary player when this player chooses strategy $s \in S$ and everybody else chooses strategy $s_0 \in S$. This new perspective on payoff functions gives rise to a natural notion of equilibrium.

**Definition 3.3 (Symmetric equilibrium).** Given a game, we say a strategy $s_e \in S$ is in symmetric equilibrium if

$$s_e \in \arg \max_{s \in S} F(s, s_e).$$

The last equilibrium concept we want to discuss, is a concept that has its roots in transportation networks [8]. In 1952 John Wardrop stated two principles which define an intuitive equilibrium concept and are applicable to transportation networks. These principles concern route choice in transportation networks and can be applied to queueing theory when we see joining one of multiple queues as a route in a transportation network. His first principle concerns a user equilibrium and his second a system optimum equilibrium.

**Definition 3.4 (User equilibrium).** Wardrop’s first principle states [8]: *The journey times on all the routes actually used are equal, and less than or equal to those which would be experienced by a single vehicle on any unused route.*

Applied to a queueing system in which a customer can join one of several queues, this means that the sojourn times of all non-empty queues are equal and less than or equal to the queues that nobody joins. This principle is called a user equilibrium as all customers individually try to minimise their own sojourn time (or travelling time in the context of transportation networks). In our case nobody has an incentive to change the queue they are standing in.

**Definition 3.5 (System optimum).** Wardrop’s second principle states [8]: *In equilibrium all users cooperate to minimise the total journey time (or equivalently the average journey time).*

In a queueing system this means that all customers cooperate to minimise the average of the expected sojourn times. Whenever a customer deviates from this system optimum, the expected sojourn time increases.

Whenever a customer is joining a queue and has to make a decision, it actually makes a difference whether the customer can see how many customers are in front of him or her. This divides queueing systems into two types: observable and unobservable queues. As the name suggests, when one joins an observable queue, one can see how many customers are in front of him waiting in line. Likewise in an unobservable queue, one can not.

As we will only regard queueing systems that are in statistical equilibrium, we will not distinguish between different customers at different times and only let customers make decisions based on the mean values in the limiting distribution. Therefore, we will not discuss any observable queues as we should make decisions of an arriving customer dependent on the number of customers that are in front of him or her upon arrival, which requires us to keep the time dependency in. However, within the class of unobservable queues we can distinguish between queues where arriving customers know the expected
sojourn time and other mean values, and queues where customers do not know these averages.

3.2. Wardrop equilibria in a priority queueing system

For this particular application we will take the priority queueing system discussed in section 2.2, but we will assume that $\mu_1 < \lambda < \mu_2$, thus that the mean service time of high priority customers is greater than the mean interarrival time of customers. That means that if everyone would join $X_1$ the system would explode. Because we want to examine this queueing model in statistical equilibrium, we need that $\rho_1 < 1$ and that $\rho_2 < 1$. From the assumption $\rho_1 < 1$ we can find the condition $q_1 < \frac{\mu_1}{\lambda}$ on this system. Also solving $\rho_2 < 1$ for $q_1$ when we assume that $\mu_1 < \mu_2$, we find a condition on $q_1$

$$q_1 < \min \left\{ \frac{\mu_1}{\lambda}, \frac{\mu_1 - \mu_2}{\lambda}, \frac{\mu_1}{\mu_1 - \mu_2} \right\},$$

which gives an upper bound on $q_1$ when this minimum is smaller than one. Of course, we should not forget that we want $q_1 \in [0, 1]$. Whenever condition (3.1) holds, the system has a statistical equilibrium distribution, and we will assume it holds throughout the remainder of this section.

We will try to find the customer equilibrium and the system optimum using Wardrop’s first and second principle, and compare the results. We will assume that customers know the expected sojourn time of both queues dependent of the fraction $q_1$, thus that the customers know $W_1(q_1)$ and $W_2(q_1)$. If we would not assume this, we can not apply Wardrop’s principles as we minimise the expected sojourn time as a function of $q_1$ in two different ways.

Throughout the remainder of this section we also assume that none of the three system parameters $\lambda$, $\mu_1$, and $\mu_2$ are equal as we will divide by the differences of two of those many times, and it will be annoying and not very satisfactory to handle these cases seperatly every time.

3.2.1. Customer equilibrium

According to Wardrop’s first principle all customers will try to minimise the expected sojourn time, individually. Whenever the expected sojourn time of $X_1$ is smaller than that of $X_2$ – so $W_1 = E[S_1] < E[S_2] = W_2$ – more customers will join $X_1$ instead of $X_2$, which increases $q_1$, the fraction of customers that join $X_1$. Likewise, when $W_2 > W_1$ more customers will join $X_2$, decreasing $q_1$. So, when $W_1 = W_2$ customers do not mind which queue they join. So, a fraction $q_1^e$ that is a solution of

$$W_1(q_1^e) = W_2(q_1^e)$$

defines a customer equilibrium. Remark that when the queue is in this equilibrium state, an arbitrary arriving customer should join $X_1$ with probability $q_1$ and $X_2$ with
probability \( q_2 = 1 - q_1 \) in order to maintain the equilibrium state – but that is exactly what we mean by \( q_1 \) and \( q_2 \).

From our discussion on M/M/1 queues in section 2.2.1 it is easily found that

\[
W_1(q_1) = \frac{1}{\mu_1 - q_1 \lambda}.
\]

(3.3)

We already found \( W_2 \) in section 2.2.2, so rewriting formula (2.8) we find

\[
W_2(q_1) = \frac{1}{\mu_2(1 - \rho_1)(1 - \rho_2)} \left( 1 + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right)
= \frac{1}{\mu_2 - \frac{\mu_2}{\mu_1} \lambda_1 - \lambda_2} + \left( \frac{\mu_1 - \lambda_2}{\mu_1 - \rho_2} \right) \left( \frac{\mu_2}{\mu_1 - \rho_1} \right)
= \frac{1}{\mu_1 - \lambda} \left( \frac{\mu_2}{\mu_1 - q_1 \lambda} + \frac{\mu_1(\lambda - \mu_1) - \mu_2(\lambda - \mu_2)}{\mu_1(\lambda - \mu_2) - q_1 \lambda(\mu_1 - \mu_2)} \right).
\]

(3.4)

Now solving the equilibrium equation (3.2) for \( q_1 \), we get

\[
\frac{\mu_1 - \mu_2 - \lambda}{\mu_1 - q_1 \lambda} = \frac{\mu_1(\lambda - \mu_1) - \mu_2(\lambda - \mu_2)}{\mu_1(\lambda - \mu_2) - q_1 \lambda(\mu_1 - \mu_2)},
\]

and it can easily be deduced that

\[
q_1^e = \frac{1}{2} \frac{\mu_1}{\lambda} \frac{\lambda + \mu_1 - \mu_2}{\mu_1 - \mu_2}.
\]

(3.5)

The next step is to find out when \( q_1^e \in [0, 1] \) and when \( q_1^e \) satisfies condition (3.1), or in other words find the conditions on \( \lambda, \mu_1, \) and \( \mu_2 \) that guarantee the existence of a (non-trivial) customer equilibrium. Because the difference in and the sum of service speed will play a central role in the following analysis, we define \( d_\mu := \mu_1 - \mu_2 \) and \( s_\mu := \mu_1 + \mu_2 \).

First we check condition (3.1),

\[
q_1^e = \frac{1}{2} \frac{\mu_1}{\lambda} \frac{\lambda + \mu_1 - \mu_2}{\mu_1 - \mu_2} < \min \left\{ \frac{\mu_1}{\lambda}, \frac{\mu_1}{\lambda} \frac{\lambda - \mu_2}{\mu_1 - \mu_2} \right\},
\]

which is satisfied when

\[
s_\mu < \lambda \text{ and } d_\mu > \lambda \text{ if } d_\mu > 0,
\]

or

\[
s_\mu > \lambda \text{ if } d_\mu < 0.
\]

Next we check whether \( q_1^e \in [0, 1] \). We see that \( q_1^e \geq 0 \) is immediately satisfied when \( d_\mu \) is positive, and that we get \( d_\mu < -\lambda \) when \( d_\mu \) is negative. The upper bound conditions depend on whether both \( 2\lambda - \mu_1 \) and \( d_\mu \) are positive or negative. If either one is negative and the other positive, then the condition is \( d_\mu \leq \frac{\mu_1 \lambda}{2\lambda - \mu_1} \), and if both are positive or negative, then we have \( d_\mu \geq \frac{\mu_1 \lambda}{2\lambda - \mu_1} \). We have summarised these conditions in table 3.1 on page 28.
Table 3.1.: Whenever these conditions depending on the signs of $d_\mu$ and $2\lambda - \mu_1$ hold, the customer equilibrium value $q_1^e$ as given by equation (3.5) can be attained.

### 3.2.2. System optimum

Wardrop’s second principle stated that in system equilibrium, or the system optimum, the average of the expected sojourn times is minimal. That is, in a transport network that is in Wardrop equilibrium the sum of the flux of all routes multiplied by the time that route takes is minimal. In our queueing system we have two routes, namely joining $X_1$ or $X_2$, each fluxes $q_1$ and $q_2 = 1 - q_1$ respectively. So the average we want to minimise is

$$W_t(q_1) := q_1W_1(q_1) + q_2W_2(q_1),$$

where $W_1$ and $W_2$ are given by equation (3.3) and (3.4) respectively. To minimise $W_t(q_1)$ we find the zeros of its derivative, as they correspond with the extrema of $W_t(q_1)$. Differentiating with respect to $q_1$ yields

$$\frac{dW_t}{dq_1}(q_1) = W_1(q_1) + q_1 \frac{dW_1}{dq_1} - W_2(q_1) + (1 - q_1) \frac{dW_2}{dq_1}. $$

The derivative of $W_1$ is easy to find,

$$\frac{dW_1}{dq_1} = \frac{\lambda}{(\mu_1 - q_1\lambda)^2},$$

and the derivative of $W_2$ is just as easy, although it looks more complex,

$$\frac{dW_2}{dq_1} = \frac{1}{\mu_1 - \lambda} \left( \frac{\mu_2\lambda}{(\mu_1 - q_1\lambda)^2} + \frac{\lambda(\mu_1 - \mu_2)(\mu_1(\lambda - \mu_1) - \mu_2(\lambda - \mu_2))}{(\mu_1(\lambda - \mu_2) - q_1\lambda(\mu_1 - \mu_2))^2} \right).$$

Plugging all these things in and simplifying the equation, we get

$$\frac{dW_t}{dq_1}(q_1) = (\mu_1 - \mu_2) \left( \frac{\lambda - \mu_1 - \mu_2}{\mu_1 - \lambda} \left( \frac{\mu_2(\lambda - \mu_1) - 1}{\mu_2(\lambda - \mu_1) - q_1^2} + \frac{1}{(\mu_1 - \lambda q_1)^2} \right) \right).$$

We want to solve $dW_t/dq_1 = 0$ in order to find the extrema of $W_t$. So we want to find $q_1^e$, the solution to

$$(\lambda - \mu_1 - \mu_2)(\mu_2(\lambda - \mu_1) - 1)(\mu_1 - \lambda q_1)^2 + (\lambda - \mu_1)(\mu_2(\lambda - \mu_1) - q_1)^2 = 0,$$
or equivalently

\[ 0 = (\lambda - \mu_1)^3 \mu_2^2 + \mu_1^2(\mu_1 + \mu_2 - \lambda)(1 + \mu_2(\mu_1 - \lambda)) \]
\[ + q_1 \left( 2\lambda\mu_1(\lambda - \mu_1 - \mu_2)(1 + \mu_2(\mu_1 - \lambda)) - 2(\lambda - \mu_1)^2 \mu_2 \right) \]
\[ + q_1^2 \left( \lambda - \mu_1 - \lambda^2(\lambda - \mu_1 - \mu_2)(1 + \mu_2(\mu_1 - \lambda)) \right). \]

It is clear that the zeros of this quadratic polynomial in \( q_1 \) are easy to determine whenever the system parameters \( \lambda, \mu_1, \) and \( \mu_2 \) are known. We will not use the quadratic formula to explicitly compute these zeros in terms of the system parameters, as it will not give us more insight in the behaviour of the zeros, and thus the extrema of \( W_t(q_1) \).

Obviously, the values found using this method might be maxima as well and might not even lie in the interval \([0, 1]\). Because we could not find a nice, explicit expression for the absolute minima we can not find conditions on the system parameters that guarantee existence of an absolute minimum in the interval \([0, 1]\) – as we did with the customer equilibrium in section 3.2.1. Also, as we restrict \( W_t(q_1) \) to the interval \([0, 1]\), there might be a minimum on the boundary \( q_1^e = 0 \) or \( q_1^e = 1 \).

### 3.3. Including service costs in a priority queueing system

Like before, we study the priority queueing model discussed in section 2.2. In section 3.2.1 we discussed an equilibrium situation for this model in which there was no incentive to choose \( X_1 \) over \( X_2 \) and vice versa. We found an equilibrium value \( q_1^e \) and set conditions on the system parameters for this value to make sense. We can, however, add system parameters such that we always have a customer equilibrium. We can do so by setting costs to joining \( X_1 \) and \( X_2 \). Service of type 1 costs \( p_1 \) in units of money (whatever they may be) and likewise service of type 2 costs \( p_2 \) units of money.

The customer equilibrium discussed before was based on the notion that every customer minimises their own travel time. We can expand this notion, incorporating that every customer tries to minimise their travel costs, where we see both money and time as travel costs. Because we need to compare time and money, we have to assume there is a (bijective) relation between the valuation of time and money – and we will further assume that this relation is the same for every customer.

For now, we assume that one unit of money is worth just as much as one unit of time – then the proverbial equality “money = time” holds, thus their relation is linear. Later, we can always incorporate a different relation between time and money by simple replacement of \( p_i \) by \( f(p_i) \), where \( f \) is our bijective relation between money and time.

The system is in equilibrium whenever a customer arrives and it does not make a difference in expected costs (of time and money) whether he or she joins \( X_1 \) or \( X_2 \). So, in equilibrium

\[ W_1(q_1) + p_1 = W_2(q_1) + p_2, \quad (3.7) \]
where $p_i$ is the price one has to pay to receive service of type $i$. We immediately see that the only thing that is dependent on the fraction $q_1$, is the difference in prices $p_1 - p_2$ and not the actual values for $p_1$ and $p_2$, so we can set $p_2 = 0$ for simplicity and allow $p_1 < 0$.

As we said in the introduction of this section, we can now find the price $p_1(q_1)$ that creates customer equilibrium for that particular value of $q_1$. However, another approach is to find the equilibrium fraction $q^*_1$ given a price $p_1$, thus reversing the roles, and finding a relation $q_1(p_1)$, which might not be well-defined. Just as in section 3.2 we assume that none of the system parameters are equal.

### 3.3.1. Setting a price to attain customer equilibrium

We use the equilibrium condition stated in equation (3.7) and the fact we could set $p_2 = 0$. Then plugging in the formulas for $W_1(q_1)$ and $W_2(q_1)$ we derived in section 3.2, we obtain an expression for the equilibrium price of service of type 1

$$p_1(q_1) = W_2(q_1) - W_1(q_1) = \frac{1}{\mu_1 - \lambda} \left( \frac{\lambda - \mu_1 + \mu_2}{\mu_1 - q_1 \lambda} + \frac{\mu_1(\lambda - \mu_1) - \mu_2(\lambda - \mu_2)}{\mu_1(\lambda - \mu_2) - q_1 \lambda(\mu_1 - \mu_2)} \right).$$

We do not exclude negative $p_1$, as we can imagine it means that customers receive money for joining $X_1$.

### 3.3.2. Customer equilibrium including service costs

Again we use the equilibrium condition given by equation (3.7) and assume $p_1$ is given and $p_2 = 0$. We try to invert equation (3.7) to find an expression for $q^*_1$, the fraction of customers that gives customer equilibrium. So, we want to solve

$$p_1 = W_2(q_1) - W_1(q_1) = \frac{1}{\mu_1 - \lambda} \left( \frac{\lambda - \mu_1 + \mu_2}{\mu_1 - q_1 \lambda} + \frac{\mu_1(\lambda - \mu_1) - \mu_2(\lambda - \mu_2)}{\mu_1(\lambda - \mu_2) - q_1 \lambda(\mu_1 - \mu_2)} \right)$$

for $q_1$. Or equivalently,

$$0 = \mu_1(\lambda - \mu_1)(\lambda + \mu_1 - \mu_2 + \mu_1(\lambda - \mu_2)p_1) \quad + [((\mu_1 - \mu_2)(\mu_1 - \mu_2 + \lambda(\mu_1 + \mu_2 - \lambda - 1)) + \mu_1(\mu_1 - \lambda)(\mu_1 - \mu_2 + \lambda(\lambda - \mu_2))p_1] q_1 \quad + \lambda(\lambda - \mu_1)(\mu_1 - \mu_2)p_1 q_1^2.$$

We see that the zeros of this quadratic polynomial in $q_1$ are easily computed once the system parameters $\lambda$, $\mu_1$, and $\mu_2$ are known. We will not explicitly use the quadratic formula to find the roots of this polynomial as the result will not be very enlightening.

Remark that when $p_1 = 0$ this equation simplifies to the same linear equation we found in section 3.2.1, so removing the service costs is consistent with our earlier result.

### 3.4. Maximum profit in a priority queueing system

In this section we continue our study of the priority queueing system we discussed in the previous sections. Just as in section 3.3 we charge a fee $p_1$ for entering the priority
queue $X_1$. For simplicity we assume that $\lambda < \mu_1$ and $\lambda < \mu_2$, so we do not have to worry about checking the stability condition (3.1). Suppose we are the owner of this queueing system and we can set the price $p_1$. We want to maximise our profit (per unit time), that is $\lambda_1 p_1$. We assume customers will find customer equilibrium eventually, and we want to maximise our long term profit. As we found an explicit expression for $p_1(q_1)$ it is easiest to maximise over $q_1 \in [0, 1]$, otherwise we should check for non realistic answers. So we maximise our profit per unit time,

$$
\max_{q_1 \in [0, 1]} \lambda q_1 p_1(q_1) = \max_{q_1 \in [0, 1]} \frac{q_1 \lambda}{\mu_1 - \lambda} \left( \frac{\lambda - \mu_1 + \mu_2}{\mu_1 - q_1 \lambda} + \frac{\mu_1 (\lambda - \mu_1) - \mu_2 (\lambda - \mu_2)}{\mu_1 (\lambda - \mu_2) - q_1 \lambda (\mu_1 - \mu_2)} \right),
$$

where we substituted $p_1(q_1)$ with our result from section 3.3.

However, we are not able to maximise this expression in general. The system parameters determine the behaviour of this graph so we will compute $q_1^{\max}$ and the corresponding price $p_1^{\max} := p_1(q_1^{\max})$ for a few specific examples. We have plotted the results for

- $(\lambda, \mu_1, \mu_2) = (0.90, 4.78, 0.92)$ in figure 3.1 on page 31,
- $(\lambda, \mu_1, \mu_2) = (1.27, 1.48, 2.22)$ in figure 3.2 on page 32,
- and $(\lambda, \mu_1, \mu_2) = (1.53, 1.98, 1.58)$ in figure 3.3 on page 32,

which we thought would represent the different shapes of the curves well after animated parameter plots.

![Figure 3.1.](image-url)  
Figure 3.1.: Graph of the profit per unit time as a function of $q_1$ for $\lambda = 0.90$, $\mu_1 = 4.78$, and $\mu_2 = 0.92$. There is a non-trivial maximum at $q_1^{\max} \approx 0.382$ with corresponding profit $\lambda q_1^{\max} p_1^{\max} \approx 0.986$.

We see that it is possible for non-trivial equilibria to exist, that is $q_1^{\max} < 1$, see figure 3.1. However, usually we will find that $q_1^{\max} = 1$ as then all customers join $X_1$. 

31
Figure 3.2.: Graph of the profit per unit time as a function of $q_1$ for $\lambda = 1.27$, $\mu_1 = 1.48$, and $\mu_2 = 2.22$. There is a trivial maximum at $q_1^{\text{max}} = 1$ with corresponding profit $\lambda q_1^{\text{max}} p_1^{\text{max}} \approx 42.56$.

Figure 3.3.: Graph of the profit per unit time as a function of $q_1$ for $\lambda = 1.53$, $\mu_1 = 1.98$, and $\mu_2 = 1.58$. There is a trivial maximum at $q_1^{\text{max}} = 1$ with corresponding profit $\lambda q_1^{\text{max}} p_1^{\text{max}} \approx 14.07$. 
generating the most money. Our preferred strategy does really depend on the system parameters.

In this chapter we have applied several notions of game theoretical equilibria to the priority queueing system we discussed in section 2.2. Of course, this system has its limitations in terms of game theoretical possibilities: there are only two queues and therefore only two expected sojourn times and two service costs (we could even simplify to just one) to consider. We ask ourselves whether we can extend the theory we discussed in chapter 2 to describe a *multiple level priority queueing system*: an $n$-level queueing system $X_1, X_2, \ldots, X_n$ where $X_1$ interrupts service of $X_2, \ldots, X_n$, and $X_2$ interrupts service of $X_3, \ldots, X_n$, and so on. Thus, customers in $X_n$ will only be served when the first $n - 1$ queues are empty. We will try to apply the theory discussed in chapter 2 to find the statistical equilibrium distribution of this $n$-level priority queue. A game theoretical analysis of this $n$-level queue is beyond the reach of this thesis.
4. Multiple priority levels

In this chapter we will discuss a queueing process we would like to call the \(n\)-level priority queue. Customers arrive according to a Poisson process with parameter \(\lambda\). There is only one service agent, but there are \(n \in \mathbb{N}\) different queues: \(X_1(t), X_2(t), \ldots, X_n(t)\). We denote the service time of queue \(i\) with \(T_i\) and \(T_i\) has an exponential distribution with mean \(\mu_i^{-1}\). Customers arrive according to a Poisson process with rate \(\lambda\). Every arriving customer joins queue \(i\) with probability \(q_i \in [0,1]\), thus customers arrive at \(X_i(t)\) according to a Poisson process with parameter \(\lambda_i := q_i \lambda\). Of course, we need \(\sum_{i=1}^n q_i = 1\) in order for this arrival process to be well-defined.

Customers of queue \(i\) are only served by the one operator when queues 1 to \(i-1\) are empty, and service of a customer of queue \(i\) is interrupted when a customer of higher priority arrives. When the queues of higher priority are emptied again, service can be resumed. That means that interruptions of service are preemptive-resume. In figure 4.1 a diagram of this queueing process is shown.

![Diagram of an n-level priority queue. Customers of type i have a service time that is exponentially distributed with rate \(\mu_i\).](image)

**Definition 4.1.** We define the uninterrupted traffic intensity of queue \(i\) by

\[
\rho_i := \frac{\lambda_i}{\mu_i} = q_i \frac{\lambda}{\mu_i} \quad \text{for} \ i \in \{1, 2, \ldots, n\}.
\]

Observe that this definition is not consistent with the definition in chapter 2, but it is more suitable for what is to come. These parameters will be important in our discussion concerning the different priority levels.
The first result from chapter 2 we want to recapitulate, is the mean completion time, which was a corollary from theorem 2.2 and given by equation 2.1

\[ E[C_i] = \frac{1}{\mu_i} \left( 1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k \right), \tag{4.1} \]

where \( B_{i-1} \) denotes the mean busy period of the server when only queue 1 to \( i - 1 \) are taken into account and \( C_i \) is the completion time of \( X_i(t) \). Recall from chapter 2 that \( X_i(t) \) has an equilibrium distribution when

\[ 1 > \lambda_i E[C_i] = \rho_i \left( 1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k \right), \tag{4.2} \]

Because we do not explicitly compute examples, we cannot check whether this condition is satisfied, so we will just assume that for every \( i \in \{1, \ldots, n\} \) this condition is satisfied and thus that we can speak of the equilibrium distribution of \( X_i(t) \), which we will denote with \( X_i \).

Another important result from chapter 2 we want to highlight is theorem 2.3, where a relation between the busy period of the system is given dependent on the busy period of the interruptions. From theorem 2.3 we get

\[ B_i = \frac{\lambda_i E[C_i] + B_{i-1} \sum_{k=1}^{i-1} \lambda_k}{\left( \sum_{k=1}^{i} \lambda_k \right) \left( 1 - \lambda_i E[C_i] \right)}, \tag{4.3} \]

where again \( B_i \) denotes the mean busy period of the server when only \( X_1, \ldots, X_i \) are taken into account.

**4.1. Mean completion time and busy period**

It is obvious that \( X_1 \) is again an M/M/1 queue with mean interarrival time \( \lambda_1 \) and mean service time \( \mu_1^{-1} \) as there are no queues to interrupt it. We want to refer to section 2.2.1 for a more precise discussion of this queue.

We wish to state two results from section 2.2.1 that we will need in the remainder of our discussion, namely the mean busy period \( B_1 \) and the mean completion time \( E[C_1] \). The latter is clear in this case: as there are no interruptions, the completion time \( C_1 \) is equal to the service time \( T_1 \), thus

\[ E[C_1] = E[T_1] = \frac{1}{\mu_1}. \]

We remark that the expression we found here coincides with equation 4.1 when we take the empty sum to equal zero. We have already found the busy period in equation 2.5 to equal

\[ B_1 = \frac{1}{\mu_1 - \lambda_1} = \frac{1}{\lambda_1} \frac{\rho_1}{1 - \rho_1}, \]
where $\rho_1 = \lambda_1/\mu_1$ as always. Also remark that this expression coincides with equation 4.3 when we take the empty sum to equal zero.

We will now inductively find the busy period $B_i$ and the mean completion time $E[C_i]$ for a general priority level $i \in \{1, \ldots, n\}$. We assume that all queues are in statistical equilibrium. In the following theorem we find explicit formulas for calculating the mean completion time and the mean busy period.

**Theorem 4.2.** Let $i \in \{1, \ldots, n\}$. If $X_1, \ldots, X_n$ all have a limiting distribution, then the mean completion time of $X_i$ and the mean busy period of the server when taking $X_1, \ldots, X_i$ into account are given by

$$E[C_i] = \frac{1}{\mu_i} \frac{1}{1 - \sum_{k=1}^{i-1} \rho_k}$$

and

$$B_i = \frac{1}{\left(\sum_{k=1}^{i} \lambda_k\right)} \frac{\sum_{k=1}^{i} \rho_k}{1 - \sum_{k=1}^{i} \rho_k},$$

where $\rho_k$ is the uninterrupted traffic intensity of $X_k$.

**Proof.** We will prove this theorem inductively. First we consider the base case of $i = 1$, then we consider general $1 < i \leq n$ whenever we know these results holds for $i - 1$.

The case $i = 1$: it is clear from what we showed earlier in this section that for $i = 1$ these relations hold if we take the empty sum to equal zero.

Now, consider $1 < i \leq n$. We assume that for $i - 1$ the following relation holds

$$B_{i-1} = \frac{1}{\sum_{k=1}^{i-1} \lambda_k} \frac{\sum_{k=1}^{i-1} \rho_k}{1 - \sum_{k=1}^{i-1} \rho_k}.$$\]

Combining this with equation 4.1 we get that the mean completion time equals

$$E[C_i] = \frac{1}{\mu_i} \left(1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k\right) = \frac{1}{\mu_i} \left(1 + \frac{\sum_{k=1}^{i-1} \rho_k}{1 - \sum_{k=1}^{i-1} \rho_k}\right)$$ \]

$$= \frac{1}{\mu_i} \frac{1}{1 - \sum_{k=1}^{i-1} \rho_k},$$

which completes the first part of the proof.

We now want to use equation 4.3 to compute the mean busy period $B_i$,

$$B_i = \frac{\lambda_i E[C_i] + B_{i-1} \sum_{k=1}^{i-1} \lambda_k}{\left(\sum_{k=1}^{i} \lambda_k\right) \left(1 - \lambda_i E[C_i]\right)} = \frac{\rho_i + \sum_{k=1}^{i-1} \rho_k}{\left(\sum_{k=1}^{i} \lambda_k\right) \left(1 - \sum_{k=1}^{i-1} \rho_k\right)}$$ \]

$$= \frac{1}{\left(\sum_{k=1}^{i} \lambda_k\right)} \frac{\rho_i + \sum_{k=1}^{i-1} \rho_k}{1 - \sum_{k=1}^{i-1} \rho_k} - \rho_i = \frac{1}{\left(\sum_{k=1}^{i} \lambda_k\right)} \frac{\sum_{k=1}^{i} \rho_k}{1 - \sum_{k=1}^{i-1} \rho_k},$$

which completes the proof for $1 < i \leq n$. By this inductive argument, we conclude that the theorem holds for all $i \in \{1, \ldots, n\}$.  

$\square$
Of course the previous theorem is only valid for an $n$-level priority queue when $X_1, \ldots, X_n$ have a joint statistical equilibrium distribution. It is necessary to check when they actually have a statistical equilibrium distribution.

### 4.2. Statistical equilibrium condition

Let us consider an arbitrary queue $i \in \{1, \ldots, n\}$. The interruptions come from the queues $X_1(t), \ldots, X_{i-1}(t)$. We remark that for the service of $X_i(t)$ it does not matter in which order the queues $1$ to $i-1$ are served, only that they empty out again. We define

$$Y_{i-1}(t) := \sum_{k=0}^{i-1} X_k(t),$$

which is the queueing process of all interruptions of service of $X_i(t)$. We see that $Y_{i-1}(t)$ is actually an $M/G/1$ queue of arrival rate $\lambda_{Y_{i-1}} := \sum_{k=1}^{i-1} \lambda_k = \lambda \sum_{k=1}^{i-1} q_k$ and with service time distribution

$$F_{T_{Y_{i-1}}}(t) = \sum_{k=1}^{i-1} \frac{q_k}{\sum_{m=1}^{i-1} q_m} (1 - e^{-\mu_k t}),$$

as with probability $q_k / \sum_{m=1}^{i-1} q_m$ a customer will be served at exponential service rate of $\mu_k$. But a system like this we did already study in section 1.3, so we can use our results from there and our results for an $M/G/1$ queue in general (stated in section 1.2) to find the equilibrium condition for this queueing system.

We define $\hat{q}_k := q_k / \sum_{m=1}^{i-1} q_m$. From section 1.2 we know that $Y_{i-1}(t)$ has an equilibrium distribution when

$$1 > \lambda_{Y_{i-1}} E[T_{Y_{i-1}}] = \left( \sum_{k=1}^{i-1} \frac{q_k}{\hat{q}_k \mu_k} \right) \sum_{k=1}^{i-1} \frac{q_k}{\mu_k} = \sum_{k=1}^{i-1} \rho_k. \quad (4.4)$$

It is obvious when look at the whole priority queue this way, say $Y_n(t) = \sum_{k=1}^n X_k(t)$, that the equilibrium condition becomes

$$1 > \sum_{k=1}^n \rho_k, \quad (4.5)$$

which implies that equation (4.4) is satisfied for all $i \in \{1, \ldots, n\}$ as $\rho_k \geq 0$ for all $k$.

Now suppose that $X_1, \ldots, X_{i-1}$ have an equilibrium distribution and that equation (4.4) is satisfied. From equation (4.2) and theorem 4.2, we then find that we have an equilibrium distribution for $X_i(t)$ when

$$1 > \rho_i \left( 1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k \right) = \rho_i \left( 1 + \frac{\sum_{k=1}^{i-1} \rho_k}{1 - \sum_{k=1}^{i-1} \rho_k} \right) = \frac{\rho_i}{1 - \sum_{k=1}^{i-1} \rho_k}. \quad (4.6)$$
However, this is equivalent to equation (4.4) for $Y_i(t)$

$$\frac{\rho_i}{1 - \sum_{k=1}^{i-1} \rho_k} < 1 \quad \Leftrightarrow \quad \rho_i < 1 - \sum_{k=1}^{i-1} \rho_k \quad \Leftrightarrow \quad \sum_{k=1}^{i} \rho_k < 1,$$

where the equivalence (*) is due to $\sum_{k=1}^{i-1} \rho_k < 1$.

We conclude from this discussion that there exists an equilibrium distribution whenever the condition stated in equation (4.5) is satisfied. We will now use our new view on the interruptions of $X_i$ to compute all the components needed for equation (2.4), which can be used to find the expected number of customers in $X_i$.

### 4.3. Mean performance measures

Throughout this chapter we assume that the $n$ queues have a joint equilibrium distribution, thus that the condition found in the previous section is met.

We restate equation (2.4),

$$L_i = \lambda_i E[C_i] + \frac{\lambda_i^2 E[C_i^2]}{2(1 - \lambda_i E[C_i])} + \frac{\lambda_i B_{i-1}^{(2)} \sum_{k=1}^{i-1} \lambda_k}{2(1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k)}, \quad (4.6)$$

where we define $B_{i-1}^{(2)} := E[T_{Y_{i-1}}^2]$, the expectation of the square of the busy period duration of $Y_{i-1} = X_1 + \cdots + X_{i-1}$. In the previous sections we already found $E[C_i]$ and $B_{i-1}$, so we are left to find $E[C_i^2]$ and $B_{i-1}^{(2)}$. We will find the former by using the relation given by equation (2.2), and the latter by interpreting the interruptions as an $M/G/1$ queue as described in the previous section.

Equation (2.2) states

$$E[C_i^2] = E[T_i^2] \left( 1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k \right)^2 + E[T_i] B_{i-1} \sum_{k=1}^{i-1} \lambda_k, \quad (4.7)$$

where $E[T_i] = \mu_i^{-1}$ and $E[T_i^2] = 2\mu_i^{-2}$. So, for this equation we also need $B_{i-1}^{(2)}$.

To obtain an expression for $B_{i-1}^{(2)}$ we will use equation (1.7),

$$B_{i-1}^{(2)} = \frac{E[(T_{Y_{i-1}})^2]}{(1 - \lambda Y_{i-1} E[T_{Y_{i-1}}])^2} = \frac{2 \sum_{k=1}^{i-1} \hat{q}_k \mu_k^2}{\left( 1 - \lambda \left( \sum_{k=1}^{i-1} q_k \left( \sum_{k=1}^{i-1} \hat{q}_k \mu_k \right) \right) \right)^3}, \quad (4.8)$$

where $\hat{q}_k := \frac{q_k}{\sum_{m=1}^{j} q_m}$, that is the rescaled fraction $q_k$. For convenient notation, we define $\rho_j^S = \sum_{k=1}^{j} \rho_k$, so we get

$$B_{i-1}^{(2)} = \frac{2 \sum_{k=1}^{i-1} \hat{q}_k \mu_k^2}{\left( 1 - \rho_j^S \right)^3}.$$
Plugging this into the last term of equation (4.6) and using theorem 4.2, we find
\[
\frac{\lambda_i B_{i-1}^{(2)} \sum_{k=1}^{i-1} \lambda_k}{2(1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k)} = \frac{\lambda_i \sum_{k=1}^{i-1} \frac{q_k}{\mu_k} \sum_{k=1}^{i-1} \lambda_k}{(1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k) (1 - \rho_{i-1}^S)^3},
\]
and this simplifies to
\[
\frac{\lambda_i B_{i-1}^{(2)} \sum_{k=1}^{i-1} \lambda_k}{2(1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k)} = \frac{\lambda_i \sum_{k=1}^{i-1} \frac{q_k}{\mu_k} \lambda}{(1 + \frac{\sum_{k=1}^{i-1} \rho_k}{\sum_{k=1}^{i-1} \rho_k}) (1 - \rho_{i-1}^S)^3} = \frac{\lambda_i \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k}}{(1 - \rho_{i-1}^S)^2},
\]
which is the third term of equation (4.6).

Now, we want to use our expression for $B_{i-1}^{(2)}$ to find $E[C_i^2]$.
\[
E[C_i^2] = \frac{2}{\mu_i^2} \left( 1 + \frac{\sum_{k=1}^{i-1} \rho_k}{1 - \frac{\sum_{k=1}^{i-1} \rho_k}{\sum_{k=1}^{i-1} \rho_k}} \right)^2 + \frac{2}{\mu_i} \left( 1 - \rho_{i-1}^S \right)^3 \sum_{k=1}^{i-1} \lambda_k
\]
\[
= \frac{2}{\mu_i^2} \left( 1 - \rho_{i-1}^S \right)^2 + \frac{2}{\mu_i} \left( 1 - \rho_{i-1}^S \right)^3 \sum_{k=1}^{i-1} \lambda_k.
\]
We already knew that
\[
\lambda_i E[C_i] = \frac{\rho_i}{1 - \sum_{k=1}^{i-1} \rho_k} = \frac{\rho_i}{1 - \rho_{i-1}^S},
\]
so we can use this and our result for $E[C_i^2]$ to find the second term in equation (4.6)
\[
\frac{\lambda_i^2 E[C_i^2]}{2(1 - \lambda_i E[C_i])} = \frac{1 - \frac{\rho_i}{1 - \rho_{i-1}^S}}{1 - \rho_{i-1}^S} \left( \frac{\lambda_i^2}{\mu_i^2} \left( 1 - \rho_{i-1}^S \right)^2 + \frac{\lambda_i^2 \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k}}{(1 - \rho_{i-1}^S)^3} \right)
\]
\[
= \frac{1 - \rho_{i-1}^S}{1 - \rho_{i-1}^S} \left( \frac{\rho_i^2}{(1 - \rho_{i-1}^S)^2} + \frac{\lambda_i \rho_i \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k}}{(1 - \rho_{i-1}^S)^3} \right).
\]
Finally, we put all these together to find
\[
L_i = \frac{\rho_i}{1 - \rho_{i-1}^S} + \frac{1 - \rho_{i-1}^S}{1 - \rho_{i-1}^S} \left( \frac{\rho_i^2}{(1 - \rho_{i-1}^S)^2} + \frac{\lambda_i \rho_i \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k}}{(1 - \rho_{i-1}^S)^3} \right) + \frac{\lambda_i \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k}}{(1 - \rho_{i-1}^S)^2}
\]
\[
= \frac{1}{1 - \rho_{i-1}^S} \left( \rho_i + \lambda_i \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k} \right),
\]
which is quite a simple expression. We recapitulate this important equation in the following theorem, as it is the most important result of this section.
Theorem 4.3. For an $n$-level priority queue in equilibrium with exponential service of mean $\mu_k$ for priority level $X_k$, the mean number of customers in $X_i$ is given by

$$L_i = \mathbb{E}[X_i] = \frac{1}{1 - \rho_i^S} \left( \rho_i + \frac{\lambda_i}{1 - \rho_i^S} \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k} \right), \quad i \in \{1, \ldots, n\}.$$  \hfill (4.9)

Example 4.4 ($i = 1$). We now want to check whether this expression corresponds to our earlier results, so we will compute $L_1$, which should correspond to the expected number of customers in an M/M/1 queue. For $i = 1$ we have that the sum $\sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k} = 0$, so

$$L_1 = \frac{\rho_1}{1 - \rho_1},$$

which is indeed the correct outcome for an M/M/1 queue.

Example 4.5 ($i = 2$). Just as with $i = 1$ we now check for $i = 2$, and find that

$$L_2 = \frac{1}{1 - \rho_1 - \rho_2} \left( \rho_2 + \frac{\lambda_2 \rho_1}{\mu_1} \right).$$

In section 2.2.2 we found in equation (2.7) that

$$L_2 = \frac{\rho_2}{1 - \rho_1 - \rho_2} \left( 1 + \frac{\mu_2}{\mu_1} \frac{\rho_1}{1 - \rho_1} \right),$$

which is obviously equal to our result using equation (4.9).

So far we have found the expected number of customers $L_i$ in queue $X_i$. From there it is easy to find the expected sojourn time $W_i$, the expected queueing time $W_{q,i}$, and the expected number of queueing customers $L_{q,i}$ as discussed in section 1.1.

Corollary 4.6. By Little’s Law, or theorem 1.4, we find that the expected sojourn time equals

$$W_i = \frac{L_i}{\lambda_i} = \frac{1}{1 - \rho_i^S} \left( \frac{1}{\mu_i} + \frac{\sum_{k=1}^{i-1} \rho_k}{1 - \rho_i^S} \right), \quad i \in \{1, \ldots, n\}.$$  

Corollary 4.7. Using $\mathbb{E}[C_i] = \frac{1}{\mu_i} \frac{1}{1 - \rho_i^S}$, we find that the expected queueing time is

$$W_{q,i} = W_i - \mathbb{E}[C_i] = \frac{1}{1 - \rho_i^S} \frac{\sum_{k=1}^{i} \rho_k}{1 - \rho_i^S}, \quad i \in \{1, \ldots, n\}.$$  

Corollary 4.8. Again by Little’s Law, we find that the mean number of customers in the queue equals

$$L_{q,i} = \lambda_i W_{q,i} = \frac{\lambda_i}{1 - \rho_i^S} \frac{\sum_{k=1}^{i} \rho_k}{1 - \rho_i^S}, \quad i \in \{1, \ldots, n\}.$$
It is nice that we could extend the theory of interrupted service of chapter 2 to find the mean values of arbitrary priority levels in an n-level priority queueing model. However, in hindsight there might have been an easier way than the use of the Laplace-Stieltjes transform of the completion time to build the theory on. We will now try to get the same results by using a mean value approach and the fact that a certain priority level does not see any customers of lower priority, but before that we will examine what the optimal order of service is.

4.4. Order of service

We are interested in the optimal order of service. Before we can really say anything about this, we should define what we mean by optimal. We will take optimal to be minimising the mean busy period of the server.

Recall from theorem 4.2 that the mean busy period of the whole system is given by

\[ B = \frac{\sum_{k=1}^{n} \frac{q_k}{\mu_k}}{1 - \lambda \sum_{k=1}^{n} \frac{q_k}{\mu_k}}. \]

When \( q_1, \ldots, q_n \) are given, it is fairly obvious that assigning the fastest service speed to the priority that receives the largest fraction of customers, and the second fastest service speed to the priority that receives the second largest fraction of customers, and so on. When we denote a fraction with \( q_k^{(m)} \) where the subscript \( k \) stands for the priority level and where the superscript \( m \) means that \( q_k^{(m)} > q_k^{(t)} \) for all \( t < m \). When we have the ordered sequence \( \mu^{(1)}, \ldots, \mu^{(n)} \), we state that we should assign priorities such that the sub- and superscript of all \( q \)'s and all \( \mu \)'s match. That means when we have \( q_k^{(m)} \) we assign \( \mu^{(m)} \) with priority level \( k \).

We will now show that this statement is true by switching two priority levels, say \( i \) and \( j \) with \( i < j \), thus \( X_i \) is of higher priority than \( X_j \). We denote the regular order of queues with \( X_k, k \in \{1, \ldots, n\} \), and we assume that it is in the order as described above. We denote the series of queues where we switched priorities of \( i \) and \( j \) with \( X_k^\hat{\phantom{k}} \). That means that the service time \( E[T_i] = \mu_j^{-1} \) and \( E[T_j] = \mu_i^{-1} \). We have given an overview in table 4.1.

<table>
<thead>
<tr>
<th>Priority level:</th>
<th>1</th>
<th>\cdots</th>
<th>i</th>
<th>\cdots</th>
<th>k</th>
<th>\cdots</th>
<th>j</th>
<th>\cdots</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular queues:</td>
<td>( X_1 )</td>
<td>( X_i )</td>
<td>( X_k )</td>
<td>( X_j )</td>
<td>( X_n )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switched queues:</td>
<td>( X_1 )</td>
<td>( X_i \sim X_j )</td>
<td>( X_k \sim X_k )</td>
<td>( X_j \sim X_i )</td>
<td>( X_n )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1.: The series of queues we get after switching, is denoted with \( X_k^\hat{\phantom{k}} \) as seen in this table. The notation \( X_i \sim X_j \) means that \( E[T_i] = E[T_j] \), but customers still arrive at \( X_i \) with rate \( \lambda_i \).
done when we have shown that
\[ \sum_{k=1}^{n} \frac{q_k}{\mu_k} < \sum_{k=1}^{n} \frac{q_k}{\mu_k} = \sum_{k=1}^{n} \frac{q_k}{\mu_k} + \frac{q_i}{\mu_i} + \frac{q_j}{\mu_j}, \]
which is equivalent to showing that
\[ \frac{q_i}{\mu_i} + \frac{q_j}{\mu_j} < \frac{q_i}{\mu_i} + \frac{q_j}{\mu_j} \iff q_i \left( \frac{1}{\mu_i} - \frac{1}{\mu_j} \right) < q_j \left( \frac{1}{\mu_i} - \frac{1}{\mu_j} \right). \]
It is easy to see that this inequality indeed holds when
\[ q_i < q_j \text{ and } \left( \frac{1}{\mu_i} - \frac{1}{\mu_j} > 0 \iff \mu_i < \mu_j \right) \]
or
\[ q_i > q_j \text{ and } \left( \frac{1}{\mu_i} - \frac{1}{\mu_j} < 0 \iff \mu_i > \mu_j \right), \]
which are both how we set up the optimal order. This proves that the assigning priorities in order of magnitude as described above minimises the busy period.

### 4.5. General service times

Just like we did throughout the rest of this chapter we are considering an \( n \)-level priority queueing system, that is we are considering \( X_1 \) to \( X_n \), where \( X_1 \) has the highest priority and \( X_n \) the lowest. However, one difference with the previous sections is that we now assume that the service time of every priority is generally distributed – so not exponentially. To emphasize, we assume that every priority has a service time distribution \( F_{T_i}(t) \). We denote \( \beta_i \) and \( \beta_i^{(2)} \) for the first and second moment of the service time \( T_i \). In figure 4.2 on page 42 a diagram of this queueing process is shown.

![Diagram of an n-level priority queue with general service time distributions](image.png)

Figure 4.2.: Diagram of an \( n \)-level priority queue with general service time distributions. Customers of type \( i \) have a service time \( T_i \) that has distribution \( F_{T_i} \).
We will now try to use a mean value analysis to find the mean number of customers for an \( n \)-level priority queue with general, but identical service times. We will end this section by finding the mean number of customers for a priority queue with non-identical general service times.

Furthermore, we assume that the equilibrium condition of section 4.2 is satisfied as in the derivation of this condition we did not use the fact that service time was exponentially distributed we can use it for general service time as well. Whenever this condition is met, we also have \( \rho_i < 1 \) for all \( 1 \leq i \leq n \) and we will use this property without further remark in what is to come.

### 4.5.1. Mean value analysis with identical service times

For this mean value analysis we assume all general service times have the same distribution. We need this assumption because the order in which we serve customers will make a difference in case they are not equal.

Let \( Y_i \) be the M/G/1 queue as defined in section 4.2, thus

\[
Y_i = \sum_{k=1}^{i} X_k.
\]

When we consider \( Y_i \) for certain \( i \in \{1, \ldots, n\} \), it does not matter for the number of customers in which order the customers are served as they all have the same service time distribution and we assumed service may be resumed after interruption, so we have that \( Y_i \) is an M/G/1 queue with arrival rate \( \lambda_{Y_i} = \sum_{k=1}^{i} \lambda_k \) and service time distribution \( F_T \) (this is due to \( F_{Y_i}^{Y_i}(t) = \sum_{k=1}^{i} \hat{q}_k F_T(t) = F_T(t) \), where \( \hat{q}_k \) is the rescaled fraction of customers that joins the queue \( X_k \)).

We are interested in the number of customers in queue \( X_i \) for certain \( i \). In order not having to worry about the number of customers that is in service we want to compute the mean number of queueing customers \( L_{q,i} \) from \( L_{Y,i} \). Because customers of priority \( i \) do not see customers of lower priority (that is of \( X_k \) with \( k > i \)), we see that

\[
L_{q,i} = L_{Y,i} - (\text{# customers queueing in } X_1 \text{ to } X_{i-1}),
\]  

but how can we compute the number of customers that is queueing in \( X_1 \) to \( X_{i-1} \)? Something we can compute is the expected time that a customer has to wait given that he or she does not join \( X_i \), which is the same as only looking to \( Y_i \), so

\[
E[T_{Y_i}^{Y_i} | \text{arriving customer does not join } X_i] = W_{Y_i}^{Y_i-1}.
\]

With Little’s Law we then see that the mean number of customers in \( X_1 \) to \( X_{i-1} \) must equal \( W_{Y_i}^{Y_i-1} \sum_{k=1}^{i} \lambda_k \) because customers arrive at rate \( \sum_{k=1}^{i} \lambda_k \) in the process \( Y_i \). We conclude that

\[
L_{q,i} = L_{Y,i} - W_{Y_i}^{Y_i-1} \sum_{k=1}^{i} \lambda_k = (W_{Y_i}^{Y_i} - W_{Y_{i-1}}^{Y_{i-1}}) \sum_{k=1}^{i} \lambda_k,
\]
where we used Little’s Law again. We are done when we compute the difference of the mean queueing times \( W^Y_i - W^X_i \).

One question that could be raised: why does the difference \( W^Y_i - W^X_i \) not equal the queueing time \( W_{q,i} \) of \( X_i \)? That is because the difference is computed in the mean queueing system \( Y_i \) where customers arrive with rate \( \sum_{k=1}^i \lambda_k \) and those arriving customers are not designated to a priority level. They are just served by the server with service time distribution \( F_T \). While by \( W_{q,i} \) we usually mean the expected waiting time when we know a customer joins \( X_i \), this is not visible in the summed system of \( Y_i \).

Continuing our derivation of \( L_{q,i} \), we can easily compute \( W^Y_i \) using equation (1.4),

\[
W^Y_i = \frac{\beta^{(2)}}{2} \left( \frac{\sum_{k=1}^i \lambda_k}{1 - \rho^S_i} - \frac{\sum_{k=1}^{i-1} \lambda_k}{1 - \rho^S_{i-1}} \right) \sum_{k=1}^i \lambda_k
\]

Using this result we conclude that

\[
L_{q,i} = \frac{\beta^{(2)}}{2} \left( \frac{\sum_{k=1}^i \lambda_k}{1 - \rho^S_i} - \frac{\sum_{k=1}^{i-1} \lambda_k}{1 - \rho^S_{i-1}} \right) \sum_{k=1}^i \lambda_k
= \frac{\beta^{(2)}}{2} \frac{\rho^S_i (1 - \rho^S_{i-1}) - \rho^S_{i-1} (1 - \rho^S_i)}{(1 - \rho^S_i)(1 - \rho^S_{i-1})} \sum_{k=1}^i \lambda_k
= \frac{\beta^{(2)}}{2} \frac{\lambda_i \sum_{k=1}^i \lambda_k}{1 - \rho^S_i}. \tag{4.11}
\]

This is by far the easiest way to find the \( L_{q,i} \) for general service time without using the results we have found before. However, for this approach to be easy, it was necessary to assume that all priorities have the same service time distribution, because otherwise equation (4.10) does not hold, as the order of service influences the waiting time and hence the number of customers that will be waiting.

### 4.5.2. Non-identical service times

When the service times are not identically distributed, we will use our earlier theory to find the following result.

**Theorem 4.9.** If \( X_1, \ldots, X_n \) denote an \( n \)-level priority queue with general service times; for all \( i \) the service time of \( X_i, T_i \), has distribution \( F_{T_i}(t) \). Then the mean number of customers in \( X_i \) is given by

\[
L_i = E[X_i] = \frac{1}{1 - \rho^S_{i-1}} \left( \rho_i + \frac{2^{-1} \lambda_i}{1 - \rho^S_i} \sum_{k=1}^i \lambda_k \beta^{(2)}_k \right), \tag{4.12}
\]

where \( \beta^{(2)}_k := E[T^2_k] \) denotes second moment of the service time of \( X_k \).
Proof. We will use the same techniques as in section 4.2 and section 4.3. We start by restating equation (4.6),

\[ L_i = \lambda_i E[C_i] + \frac{\lambda_i^2 E[C_i^2]}{2(1 - \lambda_i E[C_i])} + \frac{\lambda_i B_{i-1}^{(2)} \sum_{k=1}^{i-1} \lambda_k}{2(1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k)}, \]

but now we do need to find all four mean values \( E[C_i], E[C_i^2], B_{i-1}, \) and \( B_{i-1}^{(2)}. \) For convenience we denote \( \beta_k := E[T_k] \) and \( \beta_k^{(2)} := E[T_k^2] \) throughout the remainder of this chapter.

From our previous work it is now easy to find the four unknown mean values. We use equation (1.5) to find

\[ B_{i-1} = \frac{E[T^*_{i-1}]}{1 - \rho_{i-1}^S} = \frac{\sum_{k=1}^{i-1} \hat{q}_k \beta_k}{1 - \rho_{i-1}^S}, \]

where yet again \( \hat{q}_k := q_k / \sum_{m=1}^{i-1} q_m \) is the rescaled fraction of customers that joins \( X_k. \)

From there we use equation (2.1) to find the mean completion time

\[ E[C_i] = \beta_i \left( 1 + B_{i-1} \sum_{k=1}^{i-1} \lambda_k \right) = \frac{\beta_i}{1 - \rho_{i-1}^S}, \]

and from here we can pick up the line of thought of section 4.3.

Using equation (4.8) yields

\[ B_{i-1}^{(2)} = \frac{\sum_{k=1}^{i-1} \hat{q}_k \beta_k^{(2)}}{(1 - \rho_{i-1}^S)^3}, \]

and using equation (4.7) we find

\[ E[C_i^2] = \frac{\beta_i^{(2)}}{(1 - \rho_{i-1}^S)^2} + \frac{\beta_i}{(1 - \rho_{i-1}^S)^3} \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)}, \]

which completes our search for the four mean values and we can now plug them into the equation for \( L_i. \)

Using that \( \lambda_i E[C_i] = \rho_i / (1 - \rho_{i-1}^S), \) we find that

\[ L_i = \frac{\rho_i}{1 - \rho_{i-1}^S} + \frac{1}{2 \left( 1 - \rho_{i-1}^S \right)} \left( \frac{\lambda_i^2 \beta_i^{(2)}}{(1 - \rho_{i-1}^S)^2} + \frac{\lambda_i^2 \beta_i}{(1 - \rho_{i-1}^S)^3} \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)} \right) \]

\[ + \frac{1}{(1 - \rho_{i-1}^S)^3} \frac{\lambda_i \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)}}{2 \left( 1 + \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)} \right)}, \]

which can be easily reduced using the fact that we denote

\[ \sum_{k=1}^{i-1} \lambda_k \beta_k = \rho_{i-1}^S. \]
So, we find that

\[
L_i = \frac{\rho_i}{1 - \rho_i^S} + \frac{1}{2(1 - \rho_i^S)} \left( \frac{\lambda_i^2 \beta_i^{(2)}}{1 - \rho_i^S} + \frac{\lambda_i^2 \beta_i}{1 - \rho_i^S} \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)} \right) + \frac{\lambda_i}{2} \sum_{k=1}^{i-1} \lambda_k \beta_k^{(2)},
\]

and continuing our derivation, we find that

\[
L_i = \frac{\rho_i}{1 - \rho_i^S} + \frac{\lambda_i}{2(1 - \rho_i^S)} \sum_{k=1}^{i} \lambda_k \beta_k^{(2)}
\]

\[
= \frac{\rho_i}{1 - \rho_i^S} \left( \rho_i + \frac{\lambda_i}{2(1 - \rho_i^S)} \sum_{k=1}^{i} \lambda_k \beta_k^{(2)} \right),
\]

which completes the proof.

It is very nice that we have derived this result for general service times, however we would like to check that it corresponds to our earlier results for exponential service.

**Example 4.10** (Exponential service). We assume that all queues have exponential service with mean \( \beta_k = \mu_k^{-1} \) such that \( \beta_k^{(2)} = 2\mu_k^{-2} \). Plugging this into equation (4.12) gives

\[
L_i = \frac{1}{1 - \rho_i^S} \left( \rho_i + \frac{\lambda_i}{\mu_k} \sum_{k=1}^{i} \frac{\rho_k}{\mu_k} \right)
\]

\[
= \frac{1}{1 - \rho_i^S} \left( \rho_i + \frac{\rho_i^2}{1 - \rho_i^S} + \frac{\lambda_i}{\mu_k} \sum_{k=1}^{i} \frac{\rho_k}{\mu_k} \right)
\]

\[
= \frac{1}{1 - \rho_i^S} \left( \rho_i \frac{1 - \rho_i^S}{1 - \rho_i^S} + \frac{\lambda_i}{\mu_k} \sum_{k=1}^{i-1} \frac{\rho_k}{\mu_k} \right)
\]

from which it is easily seen that it corresponds to equation (4.9).

From our result for mean number of customers \( L_i \), the other mean values we have been studying are easily determined; we state them in the following corollaries.

**Corollary 4.11.** By Little’s Law we then find that mean sojourn time equals

\[
W_i = \frac{1}{1 - \rho_i^S} \left( \beta_i + \frac{2^{-1}}{1 - \rho_i^S} \sum_{k=1}^{i} \lambda_k \beta_k^{(2)} \right), \quad i \in \{1, \ldots, n\}.
\]

**Corollary 4.12.** We deduce that the mean queueing time is given by

\[
W_{q,i} = \frac{2^{-1}}{1 - \rho_i^S} \sum_{k=1}^{i} \lambda_k \beta_k^{(2)} \frac{1}{1 - \rho_i^S}, \quad i \in \{1, \ldots, n\}.
\]
Corollary 4.13. Finally, we find that the mean number of customers in the actual queue equals

\[ L_{q,i} = \frac{2^{-1}\lambda_i}{1 - \rho_i^S} \sum_{k=1}^{i} \alpha_{k}^{(2)} \frac{\lambda_i \beta_k^{(2)}}{1 - \rho_i^S}, \quad i \in \{1, \ldots, n\}. \]

We would also like to verify our result corresponds to the result we got using the mean value analysis given that the service times were identically distributed.

Example 4.14 (Identically distributed service times). When all service times are identically distributed – thus \( \beta_k^{(2)} = \beta^{(2)} \) for all \( k \) – it is obvious that

\[ L_{q,i} = \frac{2^{-1}\lambda_i}{1 - \rho_i^S} \sum_{k=1}^{i} \alpha_{k}^{(2)} \frac{\lambda_i \beta_k^{(2)}}{1 - \rho_i^S} = \frac{\beta^{(2)}}{2} \frac{\lambda_i}{1 - \rho_i^S} \sum_{k=1}^{i} \lambda_k, \]

from which the right hand side is the same as equation (4.11), the result we derived using the mean value analysis in section 4.5.1.

With these corollaries we have computed the mean values for an \( n \)-level priority queue with general service times. We have used the theory of chapter 1 as well as chapter 2 to derive these results. Checking with the literature ([7]) indicates that we have derived the right results. We will now try to use these results to find the so-called heavy traffic approximation for a priority queueing model.
5. Heavy traffic approximation

When the load on a queueing system increases – which might happen due to an increasing customer arrival rate or a decreasing service speed – there will be more customers waiting for service in the queue. We generally need the traffic intensity $\rho$ to be strictly smaller than unity for a steady state to exist. One might wonder what happens when we let the traffic intensity $\rho$ be near unity, but strictly smaller. In that case, it is clear that the mean queueing time $W_q$ is very large. When a queueing system has a traffic intensity that is barely smaller than unity, we say that the system is in heavy traffic, [3], [2].

In this chapter we want to study the behaviour of the $n$-level preemptive-priority queueing system with general service times in heavy traffic. Inspired by the work of Kingman$^1$, [3], we will investigate the behaviour of the mean values $(1 - \rho_n^S)W_{q,i}$ for $i \in \{1, \ldots, n\}$ as $\rho_n^S \to 1$ from below for the preemptive-resume priority queueing model we have studied in the previous chapters. Eventually we will try to find the distribution of $(1 - \rho_n^S)T_{q,i}$ as $\rho_n^S$ goes to unity.

We use the traffic intensity $\rho_n^S$ as this is the load on the system as a whole, the $n$ priority levels together. We take the limit $\rho_n^S \uparrow 1$ by decreasing the mean arrival time of customers, thus increasing $\lambda$. From

$$\rho_n^S = \sum_{k=1}^{n} \rho_k = \lambda \sum_{k=1}^{n} q_k \beta_k,$$

we see that we get the desired result by taking the limit $\lambda \uparrow (\sum_{k=1}^{n} q_k \beta_k)^{-1}$. Throughout the rest of this chapter, this is what we mean when we denote $\rho_n^S \to 1$.

Remark 5.1. It is easy to see from taking this limit that for $i < n$ we have

$$\rho_i^S \to \frac{\sum_{k=1}^{i} q_k \beta_k}{\sum_{k=1}^{n} q_k \beta_k} = 1 - \frac{\sum_{k=i+1}^{n} q_k \beta_k}{\sum_{k=1}^{n} q_k \beta_k} \quad \text{as} \quad \rho_n^S \to 1,$$

which is strictly smaller than unity. We will use this in what is to come.

5.1. Mean queueing time in heavy traffic

We can easily use our results of section 4.5.2, in particular equation (4.13), to find the expected value of $(1 - \rho_n^S)T_{q,i}$. We see that

$$\mathbb{E}[(1 - \rho_n^S)T_{q,i}] = (1 - \rho_n^S)\mathbb{E}[T_{q,i}] = (1 - \rho_n^S)W_{q,i}, \quad i \in \{1, \ldots, n\},$$

$^1$The British mathematician John Kingman found in 1962 that for any G/G/1 queue the stochastic variable $(1 - \rho)T_q$ converges in distribution to an exponential distribution as $\rho \to 1$. 

48
as $\rho^S_n$ is fixed.

Intuitively, we would say that only the last priority – that is $X_n$ – will have difficulty with digesting the increase in work. As we increase the load on the whole system, every priority level will experience an increase in mean queueing time but the last priority level will see all other increasingly busy queues interrupting its own service and hence have a drastic increase in the mean queueing time.

From remark 5.1 and equation (4.13), we immediately conclude that for $1 \leq i < n$,

$$W_{q,i} \rightarrow \frac{1}{2} \sum_{k=1}^{n} q_k \beta_k \sum_{k=1}^{i} q_k \beta_k^2 \sum_{k=i+1}^{n} q_k \beta_k,$$

as $\rho^S_n \rightarrow 1$, and hence converges for all $i < n$. Therefore, the first part of our intuition seems true: although all but the last priorities will become busier, they will be able to handle the increase in work load. From this we conclude that $E [(1 - \rho^S_n)T_{q,i}]$ converges to zero, and therefore that $(1 - \rho^S_n)T_{q,i}$ can not converge in distribution to an exponential distribution (which was the case for any G/G/1 queue, see [3]) as such a distribution can not have mean zero. Because $\lim _{\rho^S_n \rightarrow 1} (1 - \rho^S_n)T_{q,i}$ is a non-negative random variable with mean zero, we even conclude that $\lim _{\rho^S_n \rightarrow 1} (1 - \rho^S_n)T_{q,i} = 0$.

For the last queue we can not just look at $W_{q,n}$ as the expression given by equation (4.13) does not converge as $\rho^S_n \rightarrow 1$. So we need to look at the expected value of $(1 - \rho^S_n)T_{q,n}$ and then using the fact that $1 - \rho^S_n$ is constant for the expectation operator $E$ we find

$$E [(1 - \rho^S_n)T_{q,n}] \rightarrow \frac{1}{2} \sum_{k=1}^{n} q_k \beta_k^2 q_n \beta_n,$$

as $\rho^S_n \rightarrow 1$. So, we see that our intuition was right again, as from this we can conclude that $W_{q,n} \rightarrow \infty$ and thus that the last priority can not handle the increase in workload.

However, in this section we merely computed the mean values of the queueing times for different priorities. We can not say anything about the distribution of the queueing times in heavy traffic, which is even more interesting. That is why we will try to find the distribution of $(1 - \rho^S_n)T_{q,i}$ as $\rho^S_n \rightarrow 1$ in the next sections.

5.2. Laplace-Stieltjes transform of the queueing time

Although the derivation of the Laplace-Stieltjes transform of the queueing time for an arbitrary priority class $i$ is very interesting, it does not offer more insight in the particular subject we want to conclude this thesis with, namely heavy traffic. However, we do need this Laplace-Stieltjes transform in order to proceed with the heavy traffic approximation. Therefore, we will state – without proof – the result, and use it in the next section. For the proof many texts can be consulted, but we used Takagi, [7].

**Theorem 5.2.** The Laplace-Stieltjes transform of the queueing time $T_{q,i}$ for priority $i$ is given by

$$\hat{T}_{q,i}(s) = \frac{(1 - \rho^S_i) \left( s + \lambda_{i-1}^S - \lambda_{i-1}^S \hat{B}_{i-1}(s) \right)}{s - \lambda_i + \lambda_i \hat{C}_i(s)},$$

49
where the Laplace-Stieltjes transform of the completion time $C_i$ is given by theorem 2.2,

$$\hat{C}_i(s) = \hat{T}_i\left(s + \lambda_{i-1}^S - \lambda_{i-1}^S \hat{B}_{i-1}(s)\right), \quad (5.1)$$

and the Laplace-Stieltjes transform of the busy period of the server taking $X_1$ to $X_i$ into account satisfies

$$\hat{B}_{i-1}(s) = \hat{T}^{Y_{i-1}}\left(s + \lambda_{i-1}^S - \lambda_{i-1}^S \hat{B}_{i-1}(s)\right) = \sum_{k=1}^{i-1} \hat{q}_k \hat{T}_k\left(s + \lambda_{i-1}^S - \lambda_{i-1}^S \hat{B}_{i-1}(s)\right), \quad (5.2)$$

where $\hat{q}_k = q_k / \sum_{m=1}^{i-1} q_m$ is the rescaled version of $q_k$ again.

### 5.3. Distribution of the queueing time in heavy traffic

For convenience, we define

$$\hat{T}^q_{q,i}(s) := \hat{T}_{q,i}[1 - \rho_n^S]s] = \frac{(1 - \rho_i^S)\left(1 - \rho_n^S\right)s + \lambda_i^S - \lambda_{i-1}^S \hat{B}_{i-1}[1 - \rho_n^S]s]}{(1 - \rho_n^S)s - \lambda_i + \lambda_i \hat{C}_i[(1 - \rho_n^S)s]}, \quad (5.3)$$

which is the Laplace-Stieltjes transform of the random variable $(1 - \rho_n^S)T_{q,i}$. Also for convenience, we define $L_0 := (\sum_{k=1}^n q_k \beta_k)^{-1}$, so for heavy traffic we take the limit $\lambda \uparrow L_0$ as discussed before. Finally, we will denote $s^\rho := (1 - \rho_n^S)s$, and we note that $s^\rho \downarrow 0$ for all $s > 0$ when we take the limit to heavy traffic.

We can directly take the limit $\rho_n^S \to 1$ of equation (5.3) by using L’Hôpital’s rule and the fact that $\hat{B}_{i-1}(0+) = 1$ and $\hat{C}_i(0+) = 1$. For L’Hôpital’s rule to hold, we need that the numerator and the denominator are differentiable on an open interval around $\lambda = L_0$. In order to check whether this holds, we will need the derivatives (to $\lambda$) of Laplace-Stieltjes transforms of the completion time and the busy period. These can be found using equation (5.1) and equation (5.2). We will find that we need to apply L’Hôpital’s rule twice and therefore we also need that the numerator and denominator are twice differentiable and that the limit of the fraction of the second derivatives exists. We will see that they are at least twice differentiable when we assume that all service times do at least have a second moment. We will see this by just taking the derivative of the Laplace-Stieltjes transforms of the busy period and the completion time.

For convenience, we adopt the common notation,

$$\partial_\lambda := \frac{\partial}{\partial \lambda} \quad \text{and} \quad \partial_\lambda^2 := \frac{\partial^2}{\partial \lambda^2}.$$  

We also define $\sigma_{i-1} := s^\rho + \lambda_i^S - \lambda_{i-1}^S \hat{B}_{i-1}(s^\rho)$. Then, using equation (5.2) we find that

$$\partial_\lambda \hat{B}_{i-1}(s^\rho) = \left(\hat{T}^{Y_{i-1}}\right)'(\sigma_{i-1}) \cdot \left(-\left(q^\beta n\right)^S s + q_i^S - q_{i-1}^S \hat{B}_{i-1}(s^\rho) - \lambda_{i-1}^S \partial_\lambda \hat{B}_{i-1}(s^\rho)\right),$$
from which we can find that

$$\lim_{\lambda \to L_0} \partial_\lambda \hat{B}_{i-1}(s^\rho) = \frac{(q\beta)^S n_s}{\lambda_i^{n_s}} \frac{\rho_i^{n_s}}{1 - \rho_i^{n_s}},$$

where we should keep in mind that $\lambda = L_0$ in the right hand side of the equation. The fact that we could just take the total derivative of $T^{Y_i-1}$ is because the distribution of the service time is independent of $\lambda$ for all priority classes, while the one of the busy period – and hence the one of the completion time – is not, obviously.

We can do the same to find $\partial^2_\lambda \hat{B}_{i-1}(s^\rho)$, and we find that

$$\partial^2_\lambda \hat{B}_{i-1}(s^\rho) = \left(\hat{T}^{Y_i-1}\right)^{''}(\sigma_{i-1}) \cdot \left( -\left( (q\beta)^S n_s + q_i^S - q_i^{S-1} \hat{B}_{i-1}(s^\rho) - \lambda_i^{S-1} \partial_\lambda \hat{B}_{i-1}(s^\rho) \right)^2 \right.\
+ \left. \left(\hat{T}^{Y_i-1}\right)^{'}(\sigma_{i-1}) \cdot \left( -2q_i^{S-1} \partial_\lambda \hat{B}_{i-1}(s^\rho) - \lambda_i^{S-1} \partial^2_\lambda \hat{B}_{i-1}(s^\rho) \right) \right),$$

from which we find that

$$\lim_{\lambda \to L_0} \partial^2_\lambda \hat{B}_{i-1}(s^\rho) = \left[ (q\beta)^S n_s \right]^2 \frac{\sum_{k=1}^{i-1} q_k \beta^{S-1}_{k}}{(1 - \rho_i^{S-1})^3} + \frac{(q\beta)^S n_s}{\lambda_i^{n_s}} \frac{2(q\beta)^S n_s \rho_i^{n_s}}{(1 - \rho_i^{n_s})^2},$$

where yet again we should keep in mind that $\lambda = L_0$.

We will do the same for $\hat{C}_i(s^\rho)$ using equation (5.1), and we find that

$$\partial_\lambda \hat{C}_i(s) = \hat{T}_i^{'}(\sigma_{i-1}) \cdot \left( -\left( (q\beta)^S n_s + q_i^S - q_i^{S-1} \hat{B}_{i-1}(s^\rho) - \lambda_i^{S-1} \partial_\lambda \hat{B}_{i-1}(s^\rho) \right)^2 \right.\
+ \left. \left(\hat{T}^{Y_i-1}\right)^{'}(\sigma_{i-1}) \cdot \left( -2q_i^{S-1} \partial_\lambda \hat{B}_{i-1}(s^\rho) - \lambda_i^{S-1} \partial^2_\lambda \hat{B}_{i-1}(s^\rho) \right) \right),$$

from which we can conclude – using our previous results for $\hat{B}_{i-1}$ – that

$$\lim_{\lambda \to L_0} \partial_\lambda \hat{C}_i(s) = (q\beta)^S n_s \frac{\beta_i}{1 - \rho_i^{S-1}}.$$
We will now use these results to find the distribution of \((1 - \rho_n^S)T_{q,i}^\rho\) for all \(i\).

For \(i < n\) we can directly take the limit \(\rho_n^S \to 1\) of equation (5.3) by using L’Hôpital’s rule and everything we have derived so far. Then for \(s > 0\) we find

\[
\lim_{\lambda \to L_0} \hat{T}_{q,i}^\rho(s) = \lim_{\lambda \to L_0} \frac{(1 - \rho_i^S)}{1 - (1 - \rho_i^S)} \left( - (q\beta)_n^S s + q_i^S \hat{B}_{i-1} - q_i^S \hat{B}_{i-1} - \lambda_i^S \partial_\lambda \hat{C}_{i-1}(s^\rho) \right)
- \lim_{\lambda \to L_0} \frac{(q\beta)_i^S \left( (1 - \rho_i^S) s + \lambda_i^S \hat{B}_{i-1} - \lambda_i^S \hat{B}_{i-1} - \hat{C}_{i-1}(s^\rho) \right)}{1 - (1 - \rho_i^S)}
= \frac{(1 - \rho_i^S)}{1 - (1 - \rho_i^S)} \cdot \lim_{\lambda \to L_0} \frac{\lambda_i^S \cdot \partial_\lambda \hat{C}_{i-1}(s^\rho)}{\hat{C}_{i-1}(s^\rho)} = 1,
\]

where we note that after taking the limit at the equality marked with \(L\) we should keep in mind that \(\lambda = L_0\).

This result coincides with our intuition and our earlier result in section 5.1, as this means that for all \(s\) the limit

\[
\lim_{\rho_n^S \to 1} \hat{T}_{q,i}^\rho(s) = \lim_{\rho_n^S \to 1} E[e^{-s(1 - \rho_n^S)T_{q,i}}] = 1, \quad i \in \{1, \ldots, n - 1\},
\]

and thus that \((1 - \rho_n^S)T_{q,i} = 0\) almost surely with respect to the Lebesgue measure. We conclude that the priority levels 1 to \(n - 1\) can handle the increase in traffic intensity well.

We also see that we can not use this for \(i = n\) as the fraction of the derivatives to \(\lambda\) are again both zero, so we should apply L’Hôpital’s rule another time. For convenience in notation we will denote \(\hat{C}_n(s^\rho)\) just by \(\hat{C}_n\) and likewise for \(\hat{B}_{n-1}(s^\rho)\). Doing so then gives

\[
\lim_{\lambda \to L_0} \hat{T}_{q,n}^\rho(s) = \lim_{\lambda \to L_0} \frac{1}{2q_n \partial_\lambda \hat{C}_n + \lambda_n \partial_\lambda^2 \hat{C}_n} \left\{ (1 - \rho_n^S) \left( - 2q_n^S \partial_\lambda \hat{B}_{n-1} - \lambda_n^S \partial_\lambda^2 \hat{B}_{n-1} \right) \right.
- 2(q\beta)_n^S \left( - (q\beta)_n^S s + q_n^S \hat{B}_{n-1} - q_n^S \hat{B}_{n-1} - \lambda_n^S \partial_\lambda \hat{B}_{n-1} \right) \bigg\}
\leq \frac{2(q\beta)_n^S \left( q\beta)_n^S s + \lambda_n^S \hat{B}_{n-1} - \lambda_n^S \partial_\lambda \hat{B}_{n-1} \right)}{2q_n \partial_\lambda \hat{C}_n + \lambda_n \partial_\lambda^2 \hat{C}_n},
\]

where we took the limit at the equation marked with \(L\) again and from that point \(\lambda = L_0\). We immediately see from our earlier results that the numerator simplifies to

\[
(q\beta)_n^S s + \lambda_n^S \lim_{\lambda \to L_0} \partial_\lambda \hat{B}_{n-1} = \frac{(q\beta)_n^S s}{1 - \rho_n^S} = \frac{(q\beta)_n^S s}{\rho_n}.
\]
The denominator, however, is a somewhat bigger expression

\[
\lim_{\lambda \uparrow L_0} \left( 2q_n \partial_\lambda C_n + \lambda_n \partial^2_\lambda C_n \right) = 2(q\beta)_n \frac{q_n\beta_n}{\rho_n} + \left[ (q\beta)_n \sum^{n} \frac{\lambda_k \beta_k^{(2)}}{\rho_n^2} \right] + 2(q\beta)_n \frac{\rho_n^{-1} q_n^{-1}}{\rho_n \lambda_n^{-1}},
\]

which simplifies to

\[
\lim_{\lambda \uparrow L_0} \left( 2q_n \partial_\lambda C_n + \lambda_n \partial^2_\lambda C_n \right) = \left[ (q\beta)_n \sum^{n} \frac{\lambda_k \beta_k^{(2)}}{\rho_n^2} \right] + 2(q\beta)_n \frac{q_n^{-1}}{\rho_n \lambda_n^{-1}}.
\]

Plugging the previous results into the limit of the queueing time, we see that for \( s > 0 \) we get

\[
\lim_{\lambda \uparrow L_0} \hat{T}_{q,n}^o(s) = 2(q\beta)_n \frac{(q\beta)_n \sum^{n} \frac{\lambda_k \beta_k^{(2)}}{\rho_n^2}}{\rho_n} + \left[ (q\beta)_n \sum^{n} \frac{\lambda_k \beta_k^{(2)}}{\rho_n^2} \right] \frac{q_n^{-1}}{\rho_n \lambda_n^{-1}}
\]

\[
= \frac{2(q\beta)_n}{\rho_n} + \left[ (q\beta)_n \sum^{n} \frac{\lambda_k \beta_k^{(2)}}{\rho_n} \right] \frac{q_n^{-1}}{\rho_n \lambda_n^{-1}}
\]

\[
= \frac{1}{1 + s \sum^{n} \frac{q_n \beta_k^{(2)}}{q_n \beta_n}}
\]

and we recognise the Laplace-Stieltjes transform of the exponential distribution as we expected. We conclude, as \( \rho_n \to 1 \) the random variable \((1 - \rho_n)T_{q,n}\) has an exponential distribution with mean

\[
\lim_{\lambda \uparrow L_0} \mathbb{E}[1 - \rho_n T_{q,n}] = \frac{\sum^{n} \frac{q_n \beta_k^{(2)}}{q_n \beta_n}}{2q_n \beta_n},
\]

just like we found in section 5.1. This result seems in correspondence with the result of John Kingman. However, we can explicitly check it actually does correspond. In the derivation we did not choose fixed \( n \), so we can also look at the priority queue with only one priority class, thus \( n = 1 \), and this is just an M/G/1 queue for which Kingman found an explicit expression. When we consider \( n > 1 \), we know that \( X_n \) is also just an M/G/1 queue with the completion time as service time.

**Correspondence to the result of Kingman**

We will now compare our result with the result that John Kingman derived in 1962, [3]. His result is also stated in other texts, such as [2].

**Theorem 5.3** (Kingman, 1962). John Kingman found that for any G/G/1 queue the distribution of \( T_q \) is approximated by an exponential distribution with mean

\[
\mathbb{E}[T_q] \approx \frac{1}{2} \frac{\text{Var}[T] + \text{Var}[T]}{\mathbb{E}[T] - \mathbb{E}[T]},
\]

53
where $T_I$ is the interarrival time and $T$ is the service time. This is equivalent to saying that the distribution of $(1 - \rho)T_q$ goes to an exponential distribution with mean

$$\lim_{\lambda \uparrow L_0} \mathbb{E}[(1 - \rho)T_q] = \frac{1}{2} \frac{\text{Var}[T_I] + \text{Var}[T]}{\mathbb{E}[T_I]}.$$ 

By simplifying to Poisson arrivals, we can immediately compare this result with ours when we use $n = 1$, that is, simplifying our system to just an M/G/1 queue.

**Example 5.4** ($n = 1$). When we use $n = 1$ in the result of the previous section, we see that the expected value simplifies to

$$\lim_{\lambda \uparrow L_0} \mathbb{E}[(1 - \rho)T_q] = \frac{1}{2} \frac{\beta^{(2)}}{\beta},$$

where we left out the subscripts as they are unnecessary in the case $n = 1$.

In our case, we consider our priority queueing model with $n = 1$, which an M/G/1 queue, so we know that the interarrival time $T_I$ is exponentially distributed with mean $\lambda^{-1}$.

We know that the variance equals $\lambda^{-2}$. We remark that the limit we take now is $\lambda \uparrow \beta^{-1}$.

Plugging this in yields

$$\lim_{\lambda \uparrow L_0} \mathbb{E}[(1 - \rho)T_q] = \frac{1}{2} \left( \frac{1}{\lambda} + \lambda \text{Var}[T] \right) = \frac{1}{2} \left( \beta + \frac{\beta^{(2)} - \beta^2}{\beta} \right) = \frac{1}{2} \frac{\beta^{(2)}}{\beta},$$

which is exactly what we got when we used the result we derived in the previous section.

Now, for $n > 1$ we can see $X_n$ as an M/G/1 queue with the completion time as service time.

**Example 5.5** ($n > 1$). We need the completion time in this case. From section 4.5.2 we know that

$$\mathbb{E}[C_n] = \frac{\beta_n}{1 - \rho_n^S} = \frac{\beta_n}{\rho_n} = \frac{1}{\lambda_n},$$

and

$$\mathbb{E}[C_n^2] = \frac{\beta_n^{(2)}}{(1 - \rho_n^S)^2} + \frac{\beta_n}{(1 - \rho_n^S)^3} \sum_{k=1}^{n-1} \lambda_k \beta_k^{(2)} = \frac{\beta_n}{\rho_n^3} \sum_{k=1}^{n} \lambda_k \beta_k^{(2)}.$$ 

where we used that $1 - \rho_n^S = \rho_n$, which is allowed as we have already taken the limit.

Because we multiply by $1 - \rho_n^S$ but the traffic intensity $\rho$ of $X_n$ as it is used in the theorem actually equals

$$1 - \rho = 1 - \lambda_n \mathbb{E}[C_n] = \frac{1 - \rho_n^S}{1 - \rho_n^S} = \frac{1 - \rho_n^S}{\rho_n},$$

we immediately see that Kingman’s result becomes

$$\lim_{\lambda \uparrow L_0} \mathbb{E}[(1 - \rho_n^S)T_{q,n}] = \frac{\rho_n}{2} \left( \frac{1}{\lambda_n} + \lambda_n \text{Var}[C_n] \right) = \frac{\rho_n}{2} \left( \frac{1}{\lambda_n} + \lambda_n \mathbb{E}[C_n^2] - \lambda_n (\mathbb{E}[C_n])^2 \right).$$

54
Plugging the expectation values of the completion time in, we get

$$\lim_{\lambda \uparrow L_0} E[(1 - \rho_n^S) T_{q,n}] = \frac{\rho_n}{2} \left( \frac{1}{\lambda_n} + \sum_{k=1}^{n} \frac{\lambda_k \beta_k^{(2)}}{\rho_n^2} - \frac{1}{\lambda_n} \right) = \frac{1}{2} \frac{\sum_{k=1}^{n} q_k \beta_k^{(2)}}{q_n \beta_n},$$

which coincides with the result of the previous section.

We now have found that our result corresponds to the literature for all $n$.

The heavy traffic approximation

The heavy traffic results we found so far were results for the limit $\lambda \uparrow L_0$, or equivalently $\rho_n^S \to 1$. Why our results are important is because we can now approximate the queueing time when $\rho_n^S$ is very close to, but strictly smaller than unity. In that case, we see that for $t \geq 0$ we find the following approximation

$$F_{T_{q,n}}(t) = P(T_{q,n} \leq t) = P((1 - \rho_n^S) T_{q,n} \leq (1 - \rho_n^S)t) \approx 1 - \exp \left( -2 \frac{q_n \beta_n (1 - \rho_n^S)}{\sum_{k=1}^{n} q_k \beta_k^{(2)}} t \right),$$

and this final result is what we call the heavy traffic approximation.
Conclusion and review of the process

Conclusion

In this thesis, we have first examined queueing processes in general, in particular the M/G/1 queue. We extended our understanding of the M/G/1 queue to queueing processes with interrupted service, from where we could investigate a priority queueing model with preemptive-resume priority discipline and which had two priority classes: high and low. We assumed that the service times were exponentially distributed and we found formulas for the mean number of customers and the mean queueing time in terms of the system parameters. Furthermore, we discussed the stability conditions that need to hold in order for a statistical equilibrium to exist.

We also examined customer equilibrium situations, where we asked ourselves what fraction of the customers should join the high and what fraction should join the low priority queue in order for the expected waiting times to be equal. We computed the fractions for several different customer equilibria and we added the concept of paying a fee for having priority and investigated how this influenced the customer equilibria.

Then, we extended the previous model to an \( n \)-level priority queue with preemptive-resume priority discipline and general service times. We have derived formulas for the mean number of customers and mean queueing time in terms of the system parameters, and the stability conditions that need to hold to guarantee the existence of a statistical equilibrium. We found that if the sum of the traffic intensities of all uninterrupted queues is strictly smaller than unity, that is

\[
\sum_{k=1}^{n} \rho_k < 1,
\]

then the priority queue has a statistical equilibrium.

The final result of this thesis is the heavy traffic approximation, where we took the limit \( \sum_{k=1}^{n} \rho_k \to 1 \). We found that the queueing time for all but the lowest priority levels converges. For the lowest priority level we found that we can approximate the queueing time with an exponential distribution, just like Kingman [3]. We found that

\[
F_{T_{q,n}}(t) \approx 1 - e^{-\alpha(1-\rho)t},
\]

where \( T_{q,n} \) is the queueing time of the lowest priority class and \( \alpha \) is a constant that can be expressed in the system parameters.
Review of the process

At the start of this project the goal of this thesis was not completely clear. It is hard, though, how does one set a goal for himself when one does not know what his options are? What directions can I go to? I did not know. The process of studying a subject by yourself, however, is one of slow progress and growth, and slowly a direction began to dawn. When I came to him for the first time, my supervisor, prof. dr. R. Núñez-Queija, told me that I could investigate queues with processor sharing, something which is quite similar to priority queueing actually – instead of prioritising certain types of customers the server serves all customer classes at once but with different capacities. Because I did not know queueing theory that well, I simplified the model given to me by dr. Núñez-Queija to what would become section 1.3 of this thesis.

The next logical step was to add priority to the model, which I did in the next chapter. However, I was also interested in game theory and I wanted to incorporate something with the influence of customer choices as well. Finding something that really fitted the model I had made so far, was quite difficult and the final results are stated in chapter 3. Writing this chapter I found that my interest was more drawn to queueing theory, so we decided we should stay on that course – leaving chapter 3 the odd one out.

After trying to make the priority queueing system of chapter 2 more general by adapting it to arbitrary $n$ different priority levels and general service times, my supervisor suggested I could investigate the heavy traffic limit in this model. The heavy traffic limit was one I really appreciated, its simple beauty astounding me. The fact that we can approximate the queueing time in any $G/G/1$ model by an exponential distribution is in my opinion fascinating, and to derive the result myself in a way I did not find in any of the literature I have read so far, was really nice.

Of course, there is always more that can be studied. Continuing on this thesis, one could also try to find other bounds or approximations for other limiting cases than heavy traffic. It might also be interesting to model the system with computers, or try to find the optimal number of priority levels in the system (where optimal is of course free for interpretation).

I want to end this thesis by thanking dr. R. Núñez-Queija for his great supervision, his input when I did not know in which direction we should go, and the almost weekly meetings in which we discussed many results amongst other things that concerned us.
Bibliography


A. Popular summary

People deal with queueing everyday. Whether you stand in line at the supermarket to pay your groceries or, less apparently, you are waiting for your favourite internet page to load on your computer, you are actually waiting for your turn. The mathematical study of queues is called queueing theory and is a relatively young field of study – the Danish mathematician Agner Krarup Erlang was the first to write a paper on the subject in 1909.\

But what is a queue actually? A queue itself is not something we can define, or we can compute without more information. A queue is an accumulation of customers or messages (or just work) at a certain point of service. What creates a queue is nothing more than that a customer or message arrives before the previous one has been served completely. So, what defines a queue mathematically is the process according to which customers arrive and the time it takes to serve each customer, also called the service time. The arrival process is characterised by the time between the arrival of two customers, the so-called interarrival time.

Usually, we assume that both the interarrival time and the service times are stochastic of nature. That means they are certain values with certain probabilities. Let $n$ be any natural number. We denote the time between the arriving of an arbitrary customer $n$ and his successor $n + 1$ with $A_n$, thus $A_n$ is the stochastic or random variable that we call the interarrival time. We denote the service time of customer $n$ with $T_n$. Finally, we denote the probability that customer $n$ has a service time that does not exceed an arbitrary number $t > 0$ with $P(T_n \leq t)$. Likewise, we can write $P(A_n \leq a)$. It is common to assume that all customers have interarrival times and service times that are identically distributed. That means we do not have to distinguish between customers and we can just write the generic symbols (that is, without subscript) $A$ and $T$. Every time a customer arrives we draw a number from the probability distributions of $A$ and $T$, and these times are the time it takes for the next customer to come and the time it takes for the newly arrived customer to be served. This is how a queueing process works mathematically.

Now we know that the interarrival time $A$ and service time $T$ attain certain values with certain probabilities, but what value do we expect when we draw a number from their probability distributions? The expected interarrival time (also the mean interarrival time) is denoted by $E[A]$ and can be computed using an integral. Likewise, we denote the expected service time by $E[T]$. We ask ourselves, when do we expect that the queueing process is stable? That means, when do we expect that the number of customers can

---

1Agner Erlang worked for the Copenhagen Telephone Exchange and he modelled the incoming telephone calls at a telephone exchange that had to wait before being served.
become zero for an instant in finite time? It can be proven – but it is also very intuitive – that the queueing process is stable when the expected service time is strictly smaller than the expected interarrival time, that is we expect that on average less customers arrive per unit time than can be served. Mathematically we can denote the previous sentence by $E[T] < E[A]$, or

$$\frac{E[T]}{E[A]} < 1.$$  

Because the ratio in the equation above is important during the analysis of a queueing system, we give it its own name and symbol: we define the traffic intensity or workload by $\rho := \frac{E[T]}{E[A]}$ (the := means the right hand side defines the left hand side). The stability condition then becomes $\rho < 1$.

In this thesis a particular type of queueing process is studied: a priority queue that uses the preemptive-resume discipline. The idea of a priority queue is simple, there are $n$ priority classes – that is, $n$ different waiting lines all assigned with a different level of importance, or priority. Each class, for example class number $i$, has its own interarrival time denoted by the generic symbol $A_i$ and own service time denoted by $T_i$ (note that the subscript now corresponds to the class and not to the number of the customer). Although there are $n$ different queues or waiting lines, there is only one service agent that serves high priority customers first. A diagram of this queueing process is shown in figure A.1. The way the server gives priority to certain customers is stated in the discipline: preemptive-resume. Preemptive means that the server stops serving a certain customer whenever a customer of higher priority arrives. That way a customer of a certain class will only be served when there are no customers of higher priority present. Resume means that the service that has been interrupted may be resumed from where it left off after the higher priority customers have been served. This priority queueing process has been described in this thesis. We have found formulas for the mean number of customers and the mean queueing time (the time a customer actually stands in line) expressed in the system parameters.

![Diagram of an $n$-level priority queue](image)

Figure A.1.: Diagram of an $n$-level priority queue. Customers of class $i$ are served before customers of classes $i + 1, \ldots, n - 1, n$.  

60
The final part of this thesis is on the heavy traffic approximation of this priority queueing model. When a queue is in heavy traffic, the traffic intensity $\rho$ is just below unity. There are almost as many customers arriving as there can be served per unit time, and the queue will be very long – that is why it is called heavy traffic. We take the limit of $\rho \to 1$ and compute what happens with the queueing system. Intuitively we expect that only the lowest priority class experiences trouble from this critical increase in traffic as the lowest priority class can only be served when all other classes are empty. From the literature we also expect that the queueing time of the lowest priority class converges in distribution function to an exponential distribution. Both coincide with the results we found using the so called Laplace-Stieltjes transform of the queueing time.
B. Final presentation

Here is included the final presentation I gave on my thesis. My fellow mathematics students and my supervisor were in the audience.
Wie en wat?

- Rudesindo Núñez Queja
- Sindo, Markov ketens, CWI
- Onderwerp: wachtrijtheorie
- Preemptive-resume priority queue
- Afgelopen maanden hard gewerkt: af!

Wat is nu eigenlijk een queue?

- Aankomstproces van klanten of interarrival time $A$
- Service time $T$
- Stochastisch
- Vaak: aankomstproces is Poissonproces met parameter $\lambda$ of $A \sim \exp(\lambda)$
- Dit noemen we een $M/G/1$ queue
M/G/1 queue

- De queue is het proces $X(t)$: het aantal klanten op tijdstip $t$

![Diagram](image1.png)

M/G/1 queue

- De queue is het proces $X(t)$: het aantal klanten op tijdstip $t$
- Bekijk de keten $X_n := X(t_n)$ met $t_n$ tijdstip van verlaten van klant $n$
- Dan $X_n$ is een Markovketen vanwege $X_{n+1} = X_n - 1_{X_n > 0} + Y_{n+1}$ met $Y_{n+1}$ aantal aangekomen klanten tijdens de service van klant $n + 1$

![Diagram](image2.png)
M/G/1

- Op deze manier kan worden gevonden dat er een evenwichtstoestand bestaat (positief recurrent), zodra de traffic intensity
  \[ \rho := \lambda E[T] = \frac{E[T]}{E[A]} < 1 \]
- Dan kunnen we het verwachte aantal klanten uitrekenen
  \[ E[X] = \rho + \frac{\lambda^2 E[T^2]}{2(1-\rho)} \]
- Of de verwachte wachttijd tot je aan de beurt bent,
  \[ E[Q] = \frac{\lambda E[T^2]}{2(1-\rho)} \]

Priority queue

- Service times: \( T_i \sim \text{exp}(\mu_i) \)
Priority queue

- Service times: $T_i \sim \exp(\mu_i)$
- Algemene service times

Heavy traffic

- Heminer voor evenwicht moet de traffic intensity
  \[ \rho = \frac{E[T]}{E[A]} < 1 \]
- Wat nu als $\rho \to 1$?
- Wachtijd $(1 - \rho)Q_n$ voor de laagste prioriteit exponentieel verdeeld
- En voor $\rho$ dichtbij 1
  \[ P(Q_n \leq t) \approx 1 - e^{-\lambda(1-\rho)t} \]
Inhoud

- Introduction to queueing theory
- Queueing processes with interrupted service
- Customer equilibria in a priority queue
- Multiple priority levels
- Heavy traffic approximation

Bedankt voor de aandacht

Agner Krarup Erlang
Andrei Andreyevich Markov