Wachtlijnsystemen met toegewezen bedieningsstations en globale First-Come-First-Served-bediening

Queues with Dedicated Servers and Global First-Come-First-Served Scheduling

Willem Mélange

Promotoren: prof. dr. ir. H. Bruneel, prof. dr. ir. J. Walraevens
Proefschrift ingediend tot het behalen van de graad van
Doctor in de ingenieurswetenschappen

UNIVERSITEIT
GENT

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2016 - 2017

ISBN 978-90-8578-959-8
NUR 120
Wettelijk depot: D/2016/10.500/91

UNIVERSITEIT
GENT

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2016-2017

Promotoren:

Prof. Herwig Bruneel
Prof. Joris Walraevens

Leden van de examencommissie:

Prof. Gert De Cooman (voorzitter)
Prof. Chris Blondia
Prof. Dieter Claeys
Prof. Hendrik De Bie
Dr. ir. Koenraad Laevens
Prof. Sabine Wittevrongel

Dankwoord

In de eerste plaats wil ik mijn promotor Herwig Bruneel bedanken. Niet enkel omdat hij mij de kans heeft gegeven om aan dit doctoraat te werken maar ook om mij te introduceren in de wachtlijntheorie tijdens de lessen van de cursus "Wachtlijntheorie". Hij creëerde binnen SMACS een perfect kader en aangename sfeer om te werken aan mijn doctoraat waardoor het nooit echt als "werken" aanvoelde.

Daarnaast mag ik ook zeker mijn co-promotor Joris Walraevens niet vergeten. Voor het steeds klaar staan met raad en daad, het zich door mijn vele kladversies van mijn papers te ploeteren en nog zoveel meer.

Ik wil ook mijn vele collega's die van mijn vele jaren in TELIN een aangename periode gemaakt hebben, bedanken. Davy, Philippe, Sylvia en Patrick wil ik ook uitdrukkelijk bedanken voor de vele hulp bij alle technische en administratieve rompslomp en mij zo van veel frustraties te besparen.

Mijn ouders ben ik ook zeer dankbaar voor alle hulp en steun doorheen de jaren en alle mogelijkheden die ze mij geboden hebben. Zonder hen zou ik nu niet staan waar ik sta. Ook wil ik mijn broers en familie bedanken voor hun interesse en steun in mijn werk.

Tenslotte wil ik mijn vriendin Annelies bedanken omdat ze mij bij alles wat ik doe door dik en dun steunt en ook motiveert om vol te houden. Mijn prachtige dochter Arya wil ik bedanken om me met één glimlach te doen glunderen en al de rest even te doen vergeten.

Gent, augustus 2016
Willem Mélange

Table of Contents

List of Figures

List of Acronyms

E

E[X]	Expected value or mean of the random variable X

F

FCFS	First-come-first-served

G

gFCFS	Global first-come-first-served

I

IID	Independent and identically distributed

L

LST	Laplace-Stieltjes transform

M

MCF	Matrix continued fraction

P

PDF	Probability density function
PGF	Probability generating function
PMF	Probability mass function
Prob[X=n]	Probability that the random variable X is equal to n
PASTA	Poisson Arrivals See Time Averages

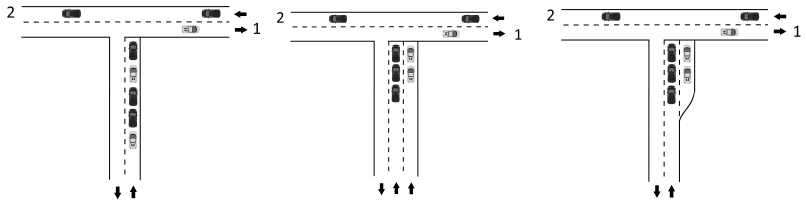
Q

QBD	Quasi-birth-death

V

Var[X]	Variance of the random variable X

Nederlandse samenvatting
–Summary in Dutch–



(a) Eén rijstrook op de hoofdbaan (b) Twee rijstroken op de hoofdbaan (c) Eén rijstrook op de hoofdbaan met een voorsorteerstrook

Figuur 1: De lichtgrijze voertuigen met bestemming 1 en de donkergrijze voertuigen met bestemming 2 naderen een kruispunt

Veelal ontstaan wachtlijnfenomenen wanneer een bepaald soort klanten, geleid door de wens om een soort van dienst te ontvangen, strijdt voor het gebruik van een bedieningsstation dat capabel is om de gevraagde dienst te leveren. Deze wachtlijnfenomenen zijn alomtegenwoordig in het alledaagse leven. Iedereen heeft reeds ervaren hoe het is om in een wachtlijn te staan. Soms kan het een aangename ervaring zijn maar meestal is het een frustrerende onderneming. Vastzitten in een file is een voorbeeld van dergelijke frustrerende onderneming en waarschijnlijk één van de meest voorkomende frustraties bij pendelaars. Het kan zelfs een negatieve invloed hebben op hun mentale gezondheid en algemeen welzijn.

Verkeerskruispunten zonder verkeerslichten zijn de meest voorkomende kruispunten. Zoals we kunnen zien in Fig. 1(a), is dergelijk kruispunt niet altijd werkconserverend en dus optimaal. Wanneer voertuigen een andere bestemming hebben maar een hoofdbaan delen, kan het zijn dat voertuigen met de éne bestemming voertuigen met een andere bestemming blokkeren. Voertuigen kunnen dus geblokkeerd zijn zelfs indien de baan richting hun bestemming vrij is. In Fig. 1 hebben alle lichtgrijze voertuigen bestemming 1 (naar rechts) en alle donkergrijze voertuigen bestemming 2 (naar links). Het donkergrijze voertuig vooraan de wachtrij wacht om zijn afslag naar links te nemen. Hierbij blokkeert dat voertuig echter het lichtgrijze voertuig achter

zich dat wel in staat is om zijn afslag naar rechts te nemen. Ideaal gezien (zoals in Fig. 1(b)), zouden de voertuigen van beide bestemmingen een eigen rijstrook op de hoofdbaan hebben. Dit is echter niet altijd fysiek mogelijk. Een mogelijke tussenoplossing is om een voorsorteerstrook te voorzien. Een voorsorteerstrook is een rijstrook met beperkte capaciteit, bestemd voor voertuigen die een bepaalde afslag willen nemen aan het volgende kruispunt (zie Fig. 1(c)). Merk op dat het blokkerende effect nog steeds mogelijk is maar minder waarschijnlijk.

Het hoofddoel van deze dissertatie is om het blokkerende effect veroorzaakt door klanten van een verschillend type (klanten die een dienst van een ander specifiek bedieningsstation nodig hebben) die een wachtrij delen, in te schatten en beter te begrijpen. Hiervoor introduceren we het concept van globale First-Come-First-Served-bediening. Alle aankomende klanten worden ondergebracht in één enkele wachtrij met specifieke bedieningsstations en worden bediend in de volgorde van hun aankomst onafhankelijk van hun type. Het tweede doel is om beter te begrijpen hoe een verslapping van de gFCFS-bediening het systeem performanter maakt. We zullen deze verslapping de gFCFS-bediening met voorsorteren noemen (P-gFCFS). Alle aankomende klanten worden ondergebracht in één enkele wachtrij met specifieke bedieningsstations en worden bediend in de volgorde van hun aankomst onafhankelijk van hun type. Dit met een uitzondering van de eerste P klanten in het wachtlijnsysteem. De eerste P klanten volgen een FCFS-bediening enkel binnen hun type. Met andere woorden, klanten met verschillende types kunnen elkaar inhalen om bediend te worden als ze bij de eerste P klanten behoren.

Hoewel we het model in zijn algemeenheid zullen bestuderen, zullen we steeds de verkeerstoepassing in het achterhoofd houden. We zullen echter ook parameters bestuderen waarover we in het verkeer geen controle hebben. Echter zullen we in ieder hoofdstuk verwijzen naar mogelijke dimensioneringsdoeleinden. Hiervoor kunnen harde en zachte beperkingen voorgesteld worden. Een harde beperking is een beperking waaraan ten alle tijde moet voldaan worden. Een zachte beperking is een beperking waaraan zo goed mogelijk wordt voldaan als de kost hiervoor niet te groot wordt. Een voorbeeld van een harde beperking is om te eisen dat de probabiliteit dat er meer dan 20 klanten in het systeem zijn kleiner is dan 10^{-5}. In de literatuur zijn reeds enkele harde beperkingen voorgesteld. Een voorsorteerstrook is noodzakelijk wanneer deze beperking niet voldaan is. Dan wordt het kruispunt namelijk als te gevaarlijk beschouwd en andere soorten kosten worden ook niet in rekening gebracht. Een heleboel andere beperkingen kunnen bedacht worden waarvan we er enkele zullen bespreken in deze dissertatie.

Deze dissertatie bestaat uit vijf hoofdstukken. In hoofdstuk 1 starten we met een korte introductie over wachtrijen en het belang van het bestuderen van wachtrijen. Daarna geven we een korte introductie over de belangrijkste concepten gebruikt in de probabiliteitstheorie en wachtlijntheorie die we ook in deze dissertatie gebruiken. Uiteindelijk bespreken we de aard van het

probleem en geven we een samenvatting van de bestaande literatuur.

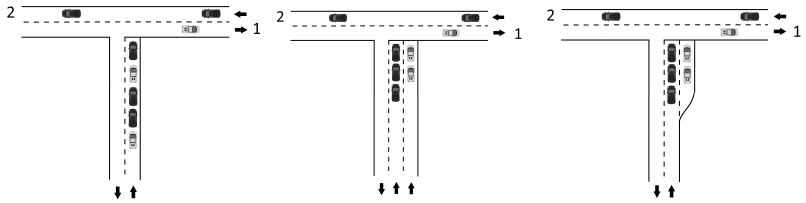
Het hoofddoel van hoofdstuk 2 is om de intuïtief verwachtte negatieve impact van de globale First-Come-First-Served-bediening te kwantificeren. Onder deze bediening verstaan we dat alle aankomende klanten worden ondergebracht in één enkele wachtrij met specifieke bedieningsstations en worden bediend in de volgorde van hun aankomst onafhankelijk van hun type. We willen analyseren of deze impact verwaarloosbaar is. Of is de impact belangrijk genoeg om in rekening te brengen. De resultaten in dit hoofdstuk geven ons reeds inzicht in het blokkerende effect. Bovendien kunnen de resultaten in dit hoofdstuk aanzien worden als een onder- en bovengrens voor het gebruik van een voorsorteerstrook (geen of een oneindig lange voorsorteerstrook). Deze grenzen kunnen ons reeds inzicht geven in de mogelijke winst van een voorsorteerstrook.

In hoofdstuk 3 introduceren we het concept van class clustering. Alle klanten van een bepaald type hebben een neiging (of geen neiging) om gegroepeerd aan te komen. Dit is een concept waarvan wij geloven dat het vaak genegeerd wordt in de literatuur maar volgens ons een grote invloed kan hebben op wachtlijnsystemen met meerdere types klanten. Het is reeds intuïtief duidelijk dat bij steeds afwisselende types klanten minder blokkering zal voorkomendan wanneer de types klanten slechts zeer zelden afwisselen. Het is dit effect dat we willen kwantificeren in dit hoofdstuk.

In hoofdstuk 4 voegen we het concept van voorsorteren toe en pakken we ons tweede objectief aan. We onderzoeken een wachtlijnmodel met een gFCFS-bediening met voorsorteren. Alle aankomende klanten worden ondergebracht in één enkele wachtrij met specifieke bedieningsstations en worden bediend in de volgorde van hun aankomst onafhankelijk van hun type. Dit met een uitzondering van de eerste P klanten in het wachtlijnsysteem. De eerste P klanten volgen een FCFS-bediening enkel binnen hun type. Met andere woorden, klanten met verschillende types kunnen elkaar inhalen om bediend te worden als ze bij de eerste P klanten behoren. De resultaten uit voorgaande hoofdstukken kunnen aanzien worden als het beste en slechtste geval. Namelijk wanneer er geen voorsorteerstrook is en wanneer er een oneindig lange voorsorteerstrook is.

Uiteindelijk zullen we in hoofdstuk 5 enkele conclusies trekken en enkele mogelijkheden voor verder onderzoek opperen.

English summary



(a) One lane on the main road (b) Two lanes on the main road (c) One lane on the main road with turn lane

Figure 2: Light grey vehicles with destination 1 and dark grey vehicles with destination 2 approaching a traffic junction

In general, queueing phenomena occur when some kind of customers, desiring to receive some kind of service, compete for the use of a service facility able to deliver the required service. These queueing phenomena are omnipresent in every day live. Everyone has experienced waiting in line. Sometimes it can be a pleasant experience, but most often it is a frustrating endeavour. Getting stuck in a traffic jam is one of these frustrating endeavours and probably one of the most common frustrations for commuters. It can even have a negative influence on their mental health and well-being.

Unsignalized intersections (crossroads and T-junctions) are the most commonly used intersections. As shown in Fig. 2(a), a junction is not always work-conserving and thus working optimally. In some cases it can even cause traffic congestion. When vehicles having different destinations share a main road, it is possible that vehicles with one destination block vehicles with another destination even though the road towards the other destination is free. In Fig. 2, all light grey vehicles have destination 1 (right) and all dark grey vehicles have destination 2 (left). The dark grey vehicle in front of the line is waiting to take its left turn. The vehicle is however also blocking the light grey vehicle behind it which could be taking a right turn. Ideally (as seen in Fig. 2(b)), both destinations would have their own lane. This is however not always physically possible. A possible workaround is a special turn lane, i.e., a lane reserved for vehicles making a specific turn at the next junction, with a smaller capacity as shown in Fig. 2(c). Notice that the blocking effect is still possible but less probable.

The main objective of this dissertation is to estimate and have a better understanding of the blocking effect caused by customers having a different type (needing service of a different dedicated server) sharing a queue. We do this by introducing the concept of the global first-come-first-served (gFCFS) restriction, i.e., all arriving customers are accommodated in one single queue with dedicated servers and are served in the order of their arrival regardless of their type. The second objective is to have a better grasp on how relaxing this gFCFS service discipline, improves the systems performance. We will call this relaxation the gFCFS with presorting service discipline, i.e., all arriving customers are accommodated in one single queue with dedicated servers and are served in the order of their arrival regardless of their type with an exception of the first P customers in the system. For the first P customers the FCFS rule holds only within the type, i.e., customers of different types can overtake each other in order to be served.

Although we will study the models generically, we will always keep the traffic application in mind. We will also study parameters of which we have no control in a traffic context. However, in each chapter we will refer to the possible dimensioning purposes. To this end, hard and soft constraints can be suggested for optimization. A hard constraint is one that must be satisfied at all times. A soft constraint is a want to be satisfied as much as possible if the cost for doing so is not too great. For example, a hard constraint could be that the probability that there are more than 20 customers in the system is less than 10^{-5}. There are already some hard constraints proposed in literature. A turn lane is warranted when the hard constraint is not fulfilled. Then the junction is considered too unsafe (prone to accidents) and other costs are not considered. However, a lot of other constraints can be imagined and some are proposed in this dissertation.

This dissertation is divided into five chapters. In Chapter 1, we start by giving a short introduction to queues and the importance of studying queue behaviour. Next, we give a short introduction to the most important concepts from probability theory and queueing theory used in this dissertation. Finally, we discuss the nature of the problem in more detail and give a literature review.

The major aim of Chapter 2 is to quantify the intuitively expected negative impact of the gFCFS service discipline, i.e., all arriving customers are accommodated in one single queue and are served in the order of their arrival regardless of their type, on the performance measures of our system. We want to analyse whether this impact is negligible or if it is important enough to take into account. The results of this chapter give us already a lot of insight into the blocking effect. Moreover, these results can also be considered in a road traffic context as lower and upper bounds for the use of a turn lane (no turn lane and an infinite turn lane). These bounds can already provide us a first insight into the potential gain of a turn lane.

In Chapter 3, we shift focus to the effect of class clustering, i.e., the way customers of any given type have a tendency to "arrive back-to-back".

Class clustering is a concept that often is neglected in literature to keep the model as simple as possible, but in this chapter we want to demonstrate that it is not always possible to treat this concept negligently. It is already intuitively clear that when the customers arrive with alternating types, less blocking will occur than when types alternate only very rarely. We quantify this effect in this chapter.

In Chapter 4, we tackle the second objective of this dissertation. This objective is to have a better grasp on the concept of the gFCFS service discipline with presorting, i.e., all arriving customers are accommodated in one single FCFS queue, regardless of their type, with an exception of the first P customers. For the first P customers the FCFS rule holds only within the type, i.e. customers of different types can overtake each other in order to be served. This models the concept of a turn lane. The result of the work in previous chapters can in fact be regarded as a worst case scenario (no turn lane), while two separate queues (infinite turn lane) can be seen as a best case scenario.

Finally, in Chapter 5 we draw some conclusions and give some possibilities for further research.

1 Introduction

1.1 Queue

In general, queueing phenomena occur when some kind of customers, desiring to receive some kind of service, compete for the use of a service facility able to deliver the required service. These queueing phenomena are omnipresent in every day life. Everyone has experienced waiting in line. Sometimes it can be a pleasant experience, but most often it is a frustrating endeavour. Below we describe some situations in which queueing is important.

Example 1.1.1. *Supermarket*
One of the most widely known queueing phenomena is probably the waiting lines in front of checkout counters in the supermarket. Everyone wants to be a "queueing expert" while searching for the fastest way to pay for their groceries. Which waiting line to pick?

The next time you have to decide which waiting line to join, keep in mind that choosing the shortest waiting line is not always the best option. Odds are that you will be waiting even longer [1].

Example 1.1.2. *Hospital Emergency Department*
This is an example where there is no actual waiting in "line". The patients wait in a waiting room and are diffused randomly. Here the patients are also no longer treated in order of arrival. Patients with more severe injuries

will get priority and will possibly get treated before other patients that are already waiting for hours.

Example 1.1.3. *Call center*
In this example customers no longer wait in a physical waiting line. Callers get placed in a virtual queue. There is no notion about the length of the queue. How many people are in front of me and how will this affect my waiting time? People are more likely to abandon the queue and retry later. This is in contrast with the previous examples where abandonment is much more unlikely.

1.2 Importance of studying queue behaviour

Understanding the nature of the queues can provide businesses a competitive advantage in the marketplace. Speed of delivery or service is being emphasized increasingly and can be partly attributed to increased competition and the value a customer places on his or her time. A customer dissatisfied by waiting in line too long for his service is a customer potentially lost and will in the worst case even share his or her bad experience with whoever will listen (negative marketing buzz). However, a satisfied customer is more likely to provide repeat business and spread the positive experience by word-of-mouth (positive marketing buzz).

Every queueing situation forms a trade-off decision between customer and system (or business) perspectives. The system will strive for an as high as possible output while minimizing costs. The customer will strive for as short as possible perceived time loss [2]. Notice here that for customer satisfaction the perceived amount of waiting time is more critical than the actual amount of waiting time [3]. It is often argued that filled time appears to pass more quickly than empty time [4].

Below we discuss the trade-off decisions and queueing experiences in the examples of Section 1.1

Example 1.2.1. *Supermarket*
The store manager will need to make the trade-off decision between the added cost of more rapid service by adding checkout counters and the inherent cost of waiting.

Example 1.2.2. *Hospital Emergency Department*
Again the trade-off decision between added cost of more rapid service by adding staff (doctors, nurses, ...) and the inherent cost of waiting, has to be made. This cost can be very high (unacceptable) if there are unnecessary casualties among the waiting patients with life-threatening injuries.

Notice that here an attempt is usually done to fill the time the patient is waiting. In waiting rooms, you can often watch television or read a magazine.

Example 1.2.3. *Call center*
Calling customer service is often done by customers feeling dissatisfied. Making this experience as pleasant as possible, is of utmost importance. Customer service can turn a dissatisfied customer into a satisfied customer. Many companies aim to answer your call, on average, in a short amount of time. A lot of effort is also put in filling your time. A lot of companies play music while you wait and give feedback about the waiting time. Sometimes they even offer to call back so that you can fill your time waiting with something useful or pleasant.

1.3 Queueing theory

Queueing theory is the scientific and mathematical field that researches queueing phenomena. A Danish engineer and mathematician A.K. Erlang working as Chief Engineer of the Copenhagen Telephone Company, is considered as the founding father of queueing theory. His work was focused on dimensioning telephone switching boards. In 1909, telephone companies still used manual telephone switchboards (Fig. 1.1). Making a call consisted of calling a telephone exchange after which the operator manually connected a cord to the proper circuit in order to complete the call. In this context, Erlang published the first "queueing" paper [5] which proves that the Poisson distribution (see later) applies to random telephone traffic. For a brief history of queueing theory in the century after this first publication, we refer to [6].

In the rest of this introductory section about queueing theory, we will discuss some important topics in the field of queueing theory (used in this dissertation). This section is intended for readers who are unfamiliar with queueing theory. For those readers who are still not satisfied after reading this section and crave for more knowledge about queueing theory, I recommend the syllabus [7] (in Dutch) used in the course "Wachtlijntheorie" by Prof. Bruneel at Ghent University where I was first introduced to the interesting field of queueing theory. The books [8] and [9] are also renowned books (in English) introducing readers to the field of queueing theory.

1.3.1 Analysis techniques

There are various techniques to study the behaviour of a queue. These techniques can be roughly divided in four categories (analytic methods, nu-

Figure 1.1: Photograph of women working at a Bell system telephone switchboard (source: U.S. National Archives)

merical methods, simulation, experimentation). The duality between these techniques lies in the contrast between simplicity and generality [10]. To keep the models tractable, analytic methods need to "strip down" the problem to its essentials. This forces one to decide what the most important parameters are (with the pitfall of neglecting crucial effects). However, analytic methods give us clear insight into the impact of certain parameters and the system in general. Experimentation on the other hand does not make any assumptions and experiments are carried out on the real system. Experimentation keeps the generality of the problem at hand but results are harder to interpret. It is often hard to determine the impact of certain parameters and whether an observation made during a run is due to a certain parameter or to the randomness built into the system. Numerical methods and simulation lie between those two extremes. Where numerical methods tend more to analytic methods (simplicity) and simulation tends more to experimentation (generality).

Generality also often brings the disadvantage of being expensive and time consuming [11]. To carry out experiments on real systems, you have to rebuild the system or interrupt day-to-day operations (losing possibly dissatisfied customers). Both are often very expensive operations. To get a true reflection of the performance of the system, it takes many weeks or even more (you can not fast-forward in real-life). These disadvantages also hold for simulation but to a lesser extent. Generality however gives a non-expert (often the one making the ultimate decision) greater confidence in the model because it is more tangible.

Notice that the techniques are complementary. First analytic or numer-

ical methods can be used to get a better insight into the problem at hand. Afterwards simulation or experimentation can be used to validate the insight (that the assumptions are not too restrictive) for the specific complex problem.

In this dissertation, we are interested in more theoretical purposes and insight in the problem at hand. We will therefore use an analytical technique, which is based on probability generating functions to solve the initial simple models. When the model becomes more complex, we will apply a numerical method based on the quasi difference equations approach.

1.3.2 Basic concepts from probability theory

1.3.2.1 Random variable

In many random experiments (every action or sequence of actions that has one or more possible outcomes), we are interested in some kind of numerical value associated with this experiment. Such variables where the exact value is dependent on the outcome of a random experiment, are called random variables (or stochastic variables).

Example 1.3.1. *Rolling a die*
An example of such random experiment is rolling a die. A common die is a small cube whose faces show numbers 1 to 6. The possible outcome can be the number on the upper surface of the die after rolling. Many random variables can be associated with this experiment. The most common random variable is the number on the upper surface itself, but also other random variables such as the square of the number on the upper surface, can be associated with the experiment.

In this dissertation, we denote random variables with capitals X, Y, ... The expected value or mean of the random variable X is denoted by $\mathrm{E}[X]$ and its variance by $\mathrm{Var}[X]$. In this dissertation, we are interested in (integer) discrete random variables (where X takes values from a finite or countable set) and continuous random variables (where the values for X form a continuum).

In probability theory, a probability distribution of the random variable X, assigns a probability to each measurable subset of possible values for X. The probability distribution of a discrete random variable can be specified by a probability mass function (pmf). This function can be defined as

$$p_X(n) \triangleq \mathrm{Prob}[X = n], n \in \mathbb{N}. \tag{1.1}$$

where $\mathrm{Prob}[X = n]$ is the probability that the random variable X is equal to n. The probability distribution of a continuous random variable can

be specified by a probability density function (pdf). This function can be defined as

$$f_X(t)dt \triangleq \text{Prob}[t < X \leq t + dt]. \tag{1.2}$$

where dt is a positive infinitesimal increment of t.

1.3.2.2 Probability generating functions

The probability generating function (pgf) $X(z)$ of a (integer) discrete random variable X is defined as

$$X(z) \triangleq \text{E}\left[z^X\right] = \sum_{n=0}^{\infty} p_X(n) z^n, \tag{1.3}$$

for all values of z for which the infinite sum converges. The pgf $X(z)$ is thus the z-transform of the pmf $p_X(n)$. Note that the pgf and pmf have a one on one relationship (given one of both, the other can be derived).

Pgfs have some very interesting properties [7], which we summarize in the remainder.

Normalization condition

$$X(1) = \sum_{n=0}^{\infty} p_X(n) = 1. \tag{1.4}$$

Analyticity and boundedness $X(z)$ is an analytic function inside the open unit disk $\{z : |z| < 1\}$ of the complex z-plane and is bounded inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z-plane. This boundedness implies among other things that the pgf does not have any poles inside or on the closed unit disk.

Moment generating property

$$\text{E}\left[\frac{X!}{(X-n)!}\right] = \left.\frac{\text{d}^n X(z)}{\text{d}z^n}\right|_{z=1}. \tag{1.5}$$

The n-th factorial moment of the random variable X can be calculated from $X(z)$ by taking the n-th derivative of $X(z)$ with respect to z and by evaluating in $z = 1$. The expected value (or mean) and variance are thus given by

$$\text{E}[X] = \left.\frac{\text{d}X(z)}{\text{d}z}\right|_{z=1}, \tag{1.6}$$

$$\text{Var}[X] = \left.\frac{\text{d}^2 X(z)}{\text{d}z^2}\right|_{z=1} + \left.\frac{\text{d}X(z)}{\text{d}z}\right|_{z=1} - \left(\left.\frac{\text{d}X(z)}{\text{d}z}\right|_{z=1}\right)^2. \tag{1.7}$$

Probability generating property

$$p_X(n) = \frac{1}{n!} \left. \frac{\mathrm{d}^n X(z)}{\mathrm{d}z^n} \right|_{z=0}. \tag{1.8}$$

In particular, this means

$$p_X(0) \triangleq \mathrm{Prob}[X = 0] = X(0). \tag{1.9}$$

1.3.2.3 Poisson distribution

A frequently used discrete distribution in this dissertation is the Poisson distribution. A Poisson random variable X with parameter λ is a discrete random variable with pmf

$$p_X(n) = e^{-\lambda} \frac{\lambda^n}{n!}, n \geq 0. \tag{1.10}$$

The related pgf $X(z)$, expected value $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$ are given by

$$X(z) = e^{\lambda(z-1)}, \tag{1.11}$$

$$\mathrm{E}[X] = \lambda, \tag{1.12}$$

$$\mathrm{Var}[X] = \lambda. \tag{1.13}$$

1.3.2.4 Laplace-Stieltjes transform

The Laplace-Stieltjes transform (LST) $X^*(s)$ of a non-negative continuous random variable X is defined as

$$X^*(s) \triangleq \mathrm{E}\left[e^{-sX}\right] = \int_0^{+\infty} f_X(t) e^{-st} \mathrm{d}t, \tag{1.14}$$

for all values of s for which the integral converges. The LST $X^*(s)$ is thus the Laplace transform of the pdf $f_X(t)$. Note that the LST and pdf have a one on one relationship (given one of both, the other can be derived).

LSTs have some interesting properties [7], which are related to the ones of pgfs.

Normalization condition

$$X^*(0) = \int_0^{\infty} f_X(t) \mathrm{d}t = 1. \tag{1.15}$$

Analyticity and boundedness $X^*(s)$ is an analytic function of s in the right half-plane of the complex s-plane ($Re(s) > 0$) and is bounded for $Re(s) \geq 0$.

Moment generating property

$$\mathrm{E}[X^n] = (-1)^n \left. \frac{\mathrm{d}^n X^*(s)}{\mathrm{d}s^n} \right|_{s=0}. \tag{1.16}$$

The n-th moment of the random variable X can be calculated from $X^*(s)$ by taking the n-th derivative of $X^*(s)$ with respect to s, giving the appropriate sign and evaluating in $s = 0$. The expected value or mean and variance are thus given by

$$\mathrm{E}[X] = - \left. \frac{\mathrm{d}X^*(s)}{\mathrm{d}s} \right|_{s=0}, \tag{1.17}$$

$$\mathrm{Var}[X] = \left. \frac{\mathrm{d}^2 X^*(s)}{\mathrm{d}s^2} \right|_{s=0} - \left(\left. \frac{\mathrm{d}X^*(s)}{\mathrm{d}s} \right|_{s=0} \right)^2. \tag{1.18}$$

1.3.2.5 Exponential distribution

A frequently used continuous distribution in this dissertation is the exponential distribution. An exponential random variable X with parameter μ is a continuous random value with the pdf

$$f_X(t) = \mu e^{-\mu t}, \mu > 0, t \geq 0. \tag{1.19}$$

The related LST $X^*(s)$, expected value $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$ are

$$X^*(s) = \frac{\mu}{\mu + s}, \quad Re(s) > -\mu, \tag{1.20}$$

$$\mathrm{E}[X] = \frac{1}{\mu}, \tag{1.21}$$

$$\mathrm{Var}[X] = \frac{1}{\mu^2}. \tag{1.22}$$

1.3.2.6 Random process

A random process (or stochastic process) $X(t)$ is an indexed collection $\{X(t), t \in I\}$ of random variables, all on the same probability space. In most engineering applications, the index set I is a set of times. If $I = \mathbb{Z}$, then $X(t)$ is called a discrete-time random process. If $I = \mathbb{R}$ or an interval of $\mathbb{R}$, then $X(t)$ is called a continuous-time random process. The value of the random variable $X(t)$ is also often referred to as the state the random process is in at time t. Some special types of random processes will be discussed next.

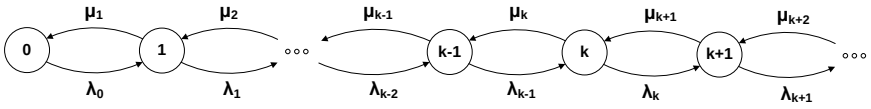


Figure 1.2: State diagram of a birth-death process

1.3.2.7 Markov process

A Markov process is a random process that satisfies the Markov property. The Markov property states that for every ordered subset $\{t_1, t_2, \cdots, t_n \mid t_1 < t_2 < \cdots < t_n\}$ of the index set I, the conditional probability distribution of $X(t_n)$ given $X(t_1), X(t_2), \cdots, X(t_{n-1})$ is equal to the conditional probability distribution of $X(t_n)$ given $X(t_{n-1})$. In words this can be loosely formulated as the "future" of the random process is only influenced by the "past" via the "present" (the way the process arrived to the "present" is irrelevant).

1.3.2.8 (Quasi-) birth-death process

Birth-death processes are a very important subclass of Markov processes. Birth-death processes are characterised by the property that only state transitions between neighbouring states occur. Going to a higher state is often referred to as a "birth", whereas going to a lower state is often referred to as a "death". The birth rate, i.e., the number of births per unit time, is denoted by $\lambda_i, i \in \mathbb{N}$ where i is the present state. The death rate, i.e., the number of deaths per unit time, is given by $\mu_i, i \in \mathbb{N}$ where i is again the present state. See Fig. 1.2 for the state diagram of a birth-death process.

A quasi-birth-death (QBD) process is a generalisation of the birth-death process. A QBD is a Markov chain (Markov process with a discrete state space) where the state space can be divided in levels with each a number of phases. The state transitions occur only between neighbouring levels or between phases of the same level. A state (m, n) is thus characterised by a level m and a phase n. See Fig. 1.3 for the state diagram of a quasi-birth-death process with two phases.

1.3.2.9 Renewal process

Renewal processes are models of stochastic phenomena in which an event occurs repeatedly over time (generally called renewals or arrivals). A stochastic process $\{N(t), t \geq 0\}$ is called a renewal process if $N(t)$ represents the total number of renewals (or arrivals) that have occurred in $(0, t]$. The inter-renewal (or inter-arrival) times $A_n = T_n - T_{n-1}$ where T_n represents

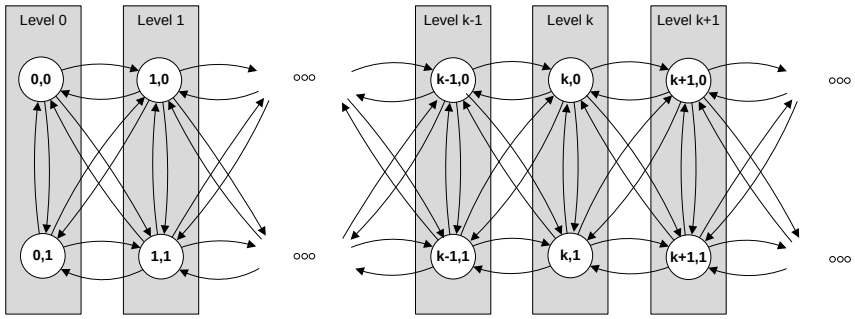


Figure 1.3: State diagram of a quasi-birth-death process with 2 phases

the time of the n-th renewal (or arrival), i.e., the time between two successive renewals (or arrivals), are independent and identically distributed (i.i.d.). Renewal processes are often found "embedded" in other stochastic processes, most notably Markov chains.

1.3.2.10 Poisson process

A Poisson process with rate λ is a special case of a renewal process where the time between successive arrivals (inter-arrival times) is exponentially distributed with parameter λ. The density of the inter-arrival times is given by

$$f_A(t) = \lambda e^{-\lambda t}, t \geq 0. \tag{1.23}$$

The total number of arrivals $N(t)$ in the interval $(0, t]$ has a Poisson distribution

$$Prob[N(t) = n] = e^{\lambda t} \frac{(\lambda t)^n}{n!}, n \geq 0. \tag{1.24}$$

In each infinitesimal time interval of length dt the occurrence of an arrival is equally likely. In other words, Poisson arrivals occur completely random in time. This is why the Poisson process is often referred to as a random arrival process. Moreover, the probability of an arrival in a certain interval is independent of the last arrival before this interval. In other words, the Poisson process does not "remember" how long ago an arrival occurred. This is why the Poisson process is often called "memoryless".

Apart from the memoryless character, Poisson processes are often used in queueing theory because of their useful properties of superposition (merging) and decomposition (splitting); Poisson processes remain Poisson processes under merging and splitting.

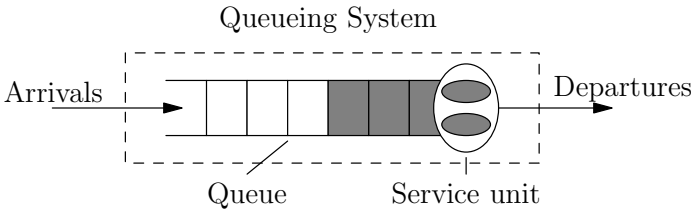


Figure 1.4: Conceptual representation of a queueing system

1.3.3 Basic concepts from queueing theory

1.3.3.1 Basic queueing model

The conceptual representation of a simple queueing system as used in this dissertation, is shown in Fig. 1.4. This can roughly be split in three parts each with their own assumptions to model the queueing system. Customers arrive at the queueing system desiring to receive some kind of service (arrivals). They wait in the queue until they receive service (queue). Afterwards, the customers exit the queueing system (departures).

The arrival process describes how customers arrive at the queueing system. In continuous-time queueing analyses, the arrival process is usually characterized by the inter-arrival times. In many practical situations, a good and simple process to describe customer arrivals is a Poisson process (see Section 1.3.2.10).

The queueing system exists of a service unit containing one or more servers who are able to offer a certain service (to one customer at a time). All servers do not necessarily offer the same service. It is possible that server 1 offers a service 1, while server 2 offers a service 2. So customers desiring to receive service 1 (2) can only be served by server 1 (2). We refer to this kind of servers as "dedicated" servers. The service time of a customer is the time the service unit needs to serve the customer. Often the service times are modelled as exponentially distributed (see Section 1.3.2.5). The service discipline determines the order in which the customers are being served. The most common service discipline is the first-come-first-served (FCFS) service discipline where customers are being served in their order of arrival.

If the arriving customer cannot be served immediately by the server, the customer has to wait in the queue. This queue or waiting line consist of a finite or infinite number of queue places (queue capacity). The capacity of a queueing system is the maximum number of customers that can be stored in the queueing system at the same time (sum of the queue capacity and number of servers).

1.3.3.2 Kendall's notation

Kendall's notation is a way to describe a queueing system. The $A \mid B \mid c$ model specifies a queueing model with c servers, the distribution of the inter-arrival times is described by A and the distribution of the service times by B. In some cases the notation is extended to 6 symbols $A \mid B \mid c \mid K \mid M \mid S$. Here K describes the capacity of the system, M the number of possible customers (population) and S the service discipline (if not mentioned either the FCFS service discipline is assumed or the service discipline is irrelevant).

Example 1.3.2. $M \mid M \mid 1$
This describes the most common queueing system in queueing theory. This describes a queueing system with one server and an infinite queue. The inter-arrival times and service times have an exponential distribution (M is short for "memoryless"). Customers are also served according to a FCFS service discipline.

Example 1.3.3. $M \mid G \mid 4 \mid 50 \mid 100 \mid LCFS$
This descriptor describes a queueing system with 4 servers (serving according to a last-come-first-served service discipline) with a finite queue having a capacity of 46. There is a population of 100 possible customers. The inter-arrival times have an exponential distribution and the service times have a general distribution.

1.3.3.3 Stability condition

Most often (as in this dissertation), a stationary analysis of the queueing model is studied. The system will reach a steady state after some transient behaviour depending on the specific initial conditions. In this steady state, the performance measures no longer change over time. For the system to be stable (or to be able to reach this steady state), the stability condition has to be fulfilled. This stability condition states that the average amount of work that enters the system per unit time should be smaller than or equal to the average amount of work the system can serve per unit time, i.e., the average amount of work the system would serve per unit time if the system was always provided with enough customers to serve. Notice that the equal sign is only possible in some special cases. If the stability condition is not fulfilled (or the system is unstable), the number of customers in a queue with infinite capacity would grow infinitely.

1.3.3.4 Performance measures

Relevant random variables in the analysis of queueing models are:

- *System occupancy* The system occupancy is defined as the number of customers in the system (including those in service).
- *System time or delay of a customer* The system time of a customer is the time a customer spends in the system (the sum of the waiting time (or wasted time) and the service time).

The most relevant and important performance measures of the queueing system are the mean values of these random values (mean system time of a customer, mean system occupancy). Other often used performance values are the tail probabilities of these random values. The tail probability is the probability that the value of a random variable is larger than a certain value.

1.3.3.5 Little's law

One of the most used and well-known laws in queueing theory is Little's law. Little's law states that, under steady-state conditions, the mean number of customers in a queueing system ($\overline{N}$) equals the mean number of arrivals in the system per unit time (λ) multiplied by the mean time spent by the customers in the system (T) [12]. Thus,

$$\overline{N} = \lambda T. \tag{1.25}$$

This law holds irrespective of the service discipline and distributions of the arrival and service processes. According to Little in [12], the intuitive reason why Little's law is true is because of a simple physical fact, i.e., "a customer in queue is also waiting". In other words, at the same time that a customer is in the system and can be counted, the customer is also accumulating minutes spent in the system. A more rigorous mathematical proof can be found in [13].

Some real-world examples where Little's Law is used in practice for some back-of-the-envelope calculations are given below.

Example 1.3.4. *Hospital Emergency Department [14]*
The length-of-stay (LOS or T in queueing terms) is becoming a key metric of focus for the emergency department. Little's law and queueing basics, give a simple interpretation of the complex relationship between emergency department staffing and LOS. For example, the Board of Directors of a hospital is demanding a LOS of 3 hours. If on average there are 30 patients in the emergency department, based on Little's law, every server in the emergency department (doctors of medicine, nurses, etc.) must process 10 patients an hour to keep up with the demand. If the doctor of medicine has average productivity of 2.5 patients an hour, four doctors of medicine are

needed to meet the demand (similar for nurses, etc.). Arrival and service variation however demand an additional staffing (generally between 10 and 20 percent).

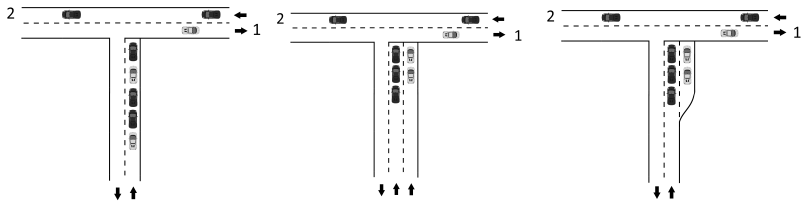
Example 1.3.5. *Manufacturing management [15]*
Little's law can help set consistent targets. Possible targets set by management are to reduce the average work-in-progress (WIP or $\overline{N}$ in queueing terms) to reduce inventory costs, to increase throughput (λ in queueing terms) or to reduce lead time (T in queueing terms) to give good customer service. For example, if a manager was told by management to achieve a three-week lead time while having a throughput of 20 jobs per week and a maximum WIP of 40 jobs, the manager would never be able to satisfy this request. If the manager would achieve a three-week lead time and a throughput of 20, the WIP would be 60. Similarly, achieving a three-week lead time and a WIP of 40 jobs would mean a throughput of 13.3 jobs a week.

1.3.3.6 PASTA property

The PASTA (Poisson Arrivals See Time Averages) property states that for queueing systems with Poisson arrivals, the fraction of arrivals that see the process in some state is equal to the fraction of time the process is in that state. Intuitively, this property holds for Poisson arrivals because Poisson arrivals occur completely randomly in time. A more rigorous mathematical proof can be found in [16].

1.4 Research question

1.4.1 The nature of the problem



(a) One lane on the main road (b) Two lanes on the main road (c) One lane on the main road with turn lane

Figure 1.5: Light grey vehicles with destination 1 and dark grey vehicles with destination 2 approaching a traffic junction

In the 2014 Annual Report, traffic monitor INRIX reveals Belgians wasted an average of 51 hours each in traffic (being number one in Europe). Getting stuck in a traffic jam is one of the most common frustrations for commuters and even has a negative influence on their mental health and well-being [17]. Traffic congestion occurs whenever the arrival rate exceeds the vehicle departure rate. This is not a new phenomenon, it is part of daily life since ancient times. It is a by-product of economic activities that grow faster than the transportation infrastructure. In this sense, congestion is a good sign. However, traffic congestion also increases air pollution and business cost (travel cost). And in the worst case, it can even have a negative impact on the economic growth when traffic congestion is too pervasive [18].

Some of the traffic congestion is recurring. Structural traffic congestion is congestion that forms every day, regardless of the weather, road accidents or other incidents. Structural traffic congestion is often caused by design and operational deficiencies. A lot of those deficiencies are linked to the history of a city. Cities typically grow in an ad-hoc manner. In ancient times most people lived within walking distance from their work, resulting in roads tending to be narrow and poorly built for motorized traffic. The vehicle fleet is also increasing and part of the traffic infrastructure is designed for a lower capacity and has now become bottlenecks.

Some of those bottlenecks where this structural traffic congestion may occur, are unsignalized intersections. Unsignalized intersections (crossroads and T-junctions) are the most commonly used intersections. As shown in Fig. 1.5(a), a junction is not always work-conserving and thus working optimally. When vehicles having different destinations share a main road, it is possible that vehicles with one destination block vehicles with another destination even though the road towards the other destination is free. In Fig. 1.5, all light grey vehicles have destination 1 (right) and all dark grey vehicles have destination 2 (left). The dark grey vehicle in front of the line is waiting to take its left turn. The vehicle is however also blocking the light grey vehicle behind it which could be taking a right turn. Ideally (as seen in Fig. 1.5(b)), both destinations would have their own lane. This is however not always physically possible. A possible workaround is a special turn lane, i.e., a lane reserved for vehicles making a specific turn at the next junction, with a smaller capacity as shown in Fig. 1.5(c). Notice that the blocking effect is still possible but less probable.

1.4.2 Research question

The main objective of this dissertation is to estimate and have a better understanding of the blocking effect caused by customers having a different type (needing service of a different dedicated server) sharing a queue. We do

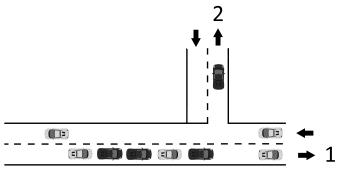


Figure 1.6: An advancing through vehicle (light grey) has been stopped behind a left-turn vehicle (dark grey)

this by introducing the concept of the global first-come-first-served (gFCFS) restriction, i.e., all arriving customers are accommodated in one single queue with dedicated servers and are served in the order of their arrival regardless of their type. The second objective is to have a better grasp on how relaxing this gFCFS service discipline, improves the systems performance. We will call this relaxation the gFCFS with presorting service discipline, i.e., all arriving customers are accommodated in one single queue with dedicated servers and are served in the order of their arrival regardless of their type with an exception of the first P customers in the system. For the first P customers the FCFS rule holds only within the type, i.e., customers of different types can overtake each other in order to be served.

Although we will study the models generically, we will always keep the traffic application in mind. We will also study parameters of which we have no control in a traffic context. However, in each chapter we will refer to the possible dimensioning purposes in a traffic context. In this dissertation, we will give a start point to decide whether or not a turn lane is worthwhile and to determine the optimal capacity for the turn lane.

1.4.3 Literature review

Harmelink was the first to publish a major contribution concerning turn lanes in [19] and provided the foundation for many current left-turn guidelines. Harmelink tried to minimize the conflict between the left-turning vehicles and through vehicles approaching from behind for safety reasons (as shown in Fig. 1.6). Using the input of traffic engineers through questionnaires, Harmelink suggested that a left-turn lane should be provided at unsignalized intersections where the probability of an advancing through vehicle that has been stopped or brought to creep-speed behind a left-turn vehicle exceeds a suitable threshold value. This suitable threshold value is dependent on the assumed operating speed. For example, for an operating speed of 40 mph (64 km/h), the maximum allowable probability of an arrival behind a left-turning vehicle is 0.02. Harmelink computes the actual probability using queueing theory, more exactly, with the $M \mid M \mid 1$ queue-

ing system and then compares this probability with the threshold value. It is well-known that the probability that there are n customers in a $M \mid M \mid 1$ queueing system (server and queue) is given by

$$\text{Prob}[N = n] = (1 - \rho)\,\rho^n, \tag{1.26}$$

where $\rho = \frac{\lambda}{\mu}$ and the probability that there are n or more customers is given by

$$\text{Prob}[N \geq n] = \rho^n. \tag{1.27}$$

The probability that there is one or more customers is thus given by ρ. The mean arrival rate is defined as the number of arrivals per hour of through vehicles behind left-turning vehicles that are waiting to make a left turn

$$\lambda = (L\,(1 - L)\,V_A)\,\frac{t_w + t_e}{\frac{2}{3}t_A}, \tag{1.28}$$

where

V_A is the advancing volume (through, left-turning and right-turning vehicles),

L is the proportion of left turns in V_A,

t_e is the average time required for a left-turning vehicle to exit from the advancing lane (based on field studies revised to 1.9 seconds),

t_A is the mean headway, i.e., the distance from the front of one vehicle to the front of the next one behind it, expressed as the time it will take for the trailing vehicle to cover that distance, in V_A $(= \frac{3600}{V_A})$

t_w is the average time that a left-turning vehicle must wait for a suitable gap in the opposing traffic stream.

Harmelink defined t_w as

$$t_w = \frac{3600}{V_O e^{-\frac{V_O G_c}{3600}}} - \frac{3600}{V_O} - G_c \tag{1.29}$$

where

V_O is the volume of traffic in the opposing lanes in vehicles per hour,

G_c is the average required headway or critical gap (determined as 5 seconds).

The mean service rate is

$$\mu = \frac{\textit{Total unblocked time}}{t_l} \tag{1.30}$$

where t_l is the average time required for making a left-turn (based on field studies revised to 3 seconds) and the total unblocked time is the time the vehicle is able to make a turn without being blocked by a vehicle of the opposite direction. Harmelink included also longer storage lengths for completeness and because some designers from his Department (Ontario Ministry of Transportation) expressed a desire that they be included. The rule of thumb is that when the storage capacity of the left-turn lane is n, then the probability of at least $n+1$ customers in the system should not exceed 0.000008 for 40 mph, 0.000003375 for 50 mph and 0.000001 for 60 mph.

Example 1.4.1. *Left-turn lane warranted and length of the lane*
Given an unsignalized intersection with assumed operating speed of 40 mph where $V_O = 400$, $V_A = 600$ of which 5% turns left. Using those parameters, $\lambda = 25.55$ and $\mu = 896$ and thus $\rho = \frac{25.55}{896} = 0.0285 > 0.02$. A left-turn lane is warranted. According to the rule of thumb, the storage capacity should be 3 since $\rho^2 = 0.0008 > 0.000008$, $\rho^3 = 0.000023 > 0.000008$ and $\rho^4 = 0.00000066 < 0.000008$.

Oppenlander and Bianchi expanded Harmelink's warrants in [20] for additional operating speeds (30 and 70 mph) and left turn percentages (ranging from 0.5% to 50%).

Fitzpatrick and Wolff reviewed several methods and state guidelines for determining when to include a left turn lane in the design at an intersection in [21]. Fitzpatrick and Wolff concluded that Harmelink's model is a widely accepted approach that is based on conflict avoidance and most of the other methods used are based on Harmelink's model. However, findings from current research suggest to revise certain assumptions made by Harmelink (G_c revised to 5.5 seconds, t_l to 4.3 seconds and t_e to 3.2 seconds). In [22], van Schalkwyk et al. also review Harmelink's model focusing on older drivers. Van Schalkwyk et al. recommended using other assumptions when larger populations of older drivers are present (G_c revised to 8 seconds, t_l to 5 seconds and t_e to 6.6 seconds).

Kikuchi and Chabroborty [23] were the first to give a critical evaluation and pointed out some limitations of Harmelink's model. They point out that there are two problems in Harmelink's formulation. The first problem is the inconsistent definitions of λ and μ. Where λ refers to the through vehicles, μ does not refer to the discharge rate of the through vehicles. This becomes critical when there are more than one through vehicles waiting behind a left-turning vehicle. The second problem is an incorrect representation of

the total number of possibilities of making a left turn in μ. Harmelink derived μ by dividing the sum of gaps that are greater than the critical gap by the time required to make a left turn (t_l). The problem in this derivation is that the residual gaps are also included in the sum of gaps and the number of opportunities to make a left-turn are exaggerated. To address these problems, Kikuchi et al. modified the equations for the arrival (λ) and service (μ) rate. These are now given by

$$\lambda^* = LV_A \left(1 - e^{\frac{-(1-L)V_A}{3600}(t_w+t_e)}\right), \tag{1.31}$$

$$\mu^* = \left(1 - e^{-3\frac{V_O}{3600}}\right) V_O \sum_{n=1}^{N} n \left(e^{-\frac{V_O}{3600}\{G_c + 3(n-1)\}}\right). \tag{1.32}$$

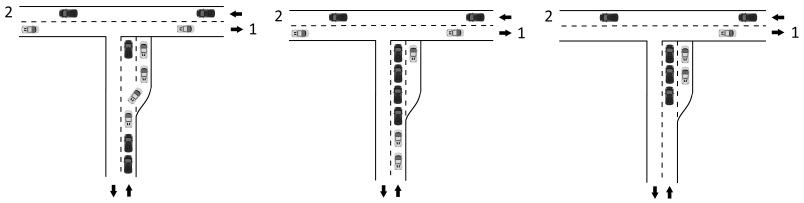
where the value of N is the maximum number of left-turning opportunities per single headway. Kikuchi and Chabroborty also proposed two extra criteria for left-turn lane warrants aside of the modified probability based criterion (Harmelink's model). Namely, a criteria based on delay to through vehicles and on the level of service. Those criteria have not been validated and did not get any continuation in further research.

Chabroborty and Kikuchi developed a model based on the $M \mid G \mid 1$ queueing system which calculates the probability that a given length of turning lane will result in overflows [24]. Lane lengths are suggested such that the probability of lane overflow is less than a given threshold value. In their paper, Chabroborty and Kikuchi validated their model by comparing the results with computer simulation software.

In [25], Lertworawanich and Elefteriadou used the $M \mid G2 \mid 1$ queueing system to determine the length of the left-turn lane. The $M \mid G2 \mid 1$ queueing system differs from the $M \mid G \mid 1$ queueing system because it considers two different types of service times. In this case, the service times for vehicles that arrive when the left-turning lane is empty differs from the service times for vehicles when the left-turning lane is not empty as discussed by Yeo and Weesakul in [26].

Notice that almost all previous papers used the lane overflow approach because of safety reasons on roads where the vehicles arrive on the major road (the through vehicles do not need to stop and the left-turning vehicles form a hazard on the road). An appropriate threshold value of lane overflow is used to warrant a left-turn lane and determine the length of the left-turn lane. However, lane blockage (as shown in Fig. 1.7(b)) can also have a significant impact on the system performance measures, especially when the proportion of left-turning vehicles is large. Notice that lane blockage is the same as lane overflow of right-turning vehicles. The possibility of lane blockage becomes much more possible when we consider the case

where vehicles arrive on a minor road (the through vehicles also need to stop). Queueing at unsignalized intersections where the vehicles arrive on the minor road has been studied comprehensively [26–29]. A study focusing on a right-turn lane at an unsignalized intersection is much more rare. Cottrell was the first to report such a study in [30] but did not use a mathematical model. Cottrell based his study on the collection of conflict data along with data on approach volume and right-turn volume. McCoy et al. developed guidelines for right-turn lanes in [31] using computer simulation. In this dissertation we give a starting point to consider specific turn lanes at unsignalized intersections when vehicles arrive on the minor road.



(a) Lane overflow (b) Lane blockage (c) No lane blockage or overflow

Figure 1.7: Light grey vehicles with destination 1 and dark grey vehicles with destination 2 approaching a traffic junction

1.5 Outline

We have divided the rest of this dissertation into three main chapters. In chapter 2, we first study the blocking effect by using a simple basic queueing model. We focus on the concept of the global first-come-first-served service discipline, i.e., all arriving customers are accommodated in one single queue and are served in the order of their arrival regardless of their type.

In chapter 3 we introduce the concept of class clustering, i.e., customers of any given type may (or may not) have a tendency to "arrive back-to-back". A concept that is often neglected in literature but which we believe has an considerable impact on the performance of multiclass queueing systems.

In chapter 4 we add the concept of presorting (the concept of a turn lane). We examine a queueing model with a gFCFS with presorting service discipline, i.e., all arriving customers are accommodated in one single queue and are served in the order of their arrival regardless of their type with an exception of the first P customers in the system. For the first P customers, the FCFS rule holds only within the type, i.e., customers of different types

can overtake each other in order to be served.

Finally, in chapter 5 we draw some conclusions.

1.6 Publications

1.6.1 Publications in journals

1. Herwig Bruneel, Willem Mélange, Bart Steyaert, Dieter Claeys and Joris Walraevens. *A Two-Class Discrete-Time Queueing Model with Two Dedicated Servers and Global FCFS Service Discipline.* European Journal of Operational Research, 223(1):123-132, 2012.

2. Herwig Bruneel, Willem Mélange, Bart Steyaert, Dieter Claeys and Joris Walraevens. *Effect of Global FCFS and Relative Load Distribution in Two-Class Queues with Dedicated Servers.* 4OR-A Quarterly Journal of Operations Research, 11(4):375-391, 2013.

3. Dieter Claeys, Herwig Bruneel, Bart Steyaert, Willem Mélange and Joris Walraevens. *Influence of Data Clustering on In-Order Multi-Core Processing Systems.* Electronics Letters, 49(1):28-29, 2013.

4. Willem Mélange, Herwig Bruneel, Bart Steyaert, Dieter Claeys and Joris Walraevens. *A Continuous-Time Queueing Model with Class Clustering and Global FCFS Service Discipline*, Journal of Industrial and Management Optimization, 10(1):193-206, 2014.

5. Willem Mélange, Joris Walraevens, Dieter Claeys, Bart Steyaert and Herwig Bruneel. *The Impact of a Global FCFS Service Discipline in a Two-Class Queue With Dedicated Servers.* Computers and Operations Research, 71:23-33, 2016.

6. Herwig Bruneel, Willem Mélange, Dieter Claeys and Joris Walraevens. *A Two-Class Global FCFS Discrete-Time Queueing Model with Arbitrary-Length Constant Service Times.* TOP, 1-15, 2016.

1.6.2 Publications in international conferences

1. Willem Mélange, Joris Walraevens, Bart Steyaert and Herwig Bruneel. *A Two-Class Queueing Model With Class Clustering and Global FCFS Service Discipline.* In Abstracts of the International Conference of Stochastic Modelling and Simulation, Chennai, India, December 15-17, page 38, 2011.

2. Willem Mélange, Herwig Bruneel, Bart Steyaert and Joris Walraevens. *A Two-Class Continuous-Time Queueing Model With Dedicated Servers and Global FCFS Service Discipline.* Lecture Notes in Computer Science (Proceedings of the ASMTA 2011 Conference), Venezia, Italy, June 20-22, 6751:14-27, 2011.

3. Herwig Bruneel, Willem Mélange, Bart Steyaert, Dieter Claeys and Joris Walraevens. *Influence of Relative Traffic Distribution in Nodes With Blocking: an Analytical Model.* In European Simulation and Modelling Conference 2012, Essen, Germany, November 22-24, pages 136-143, 2012.

4. Willem Mélange, Herwig Bruneel, Dieter Claeys, Bart Steyaert and Joris Walraevens. *Impact of Class Clustering and Global FCFS Service Discipline on the System Occupancy of a Two-Class Queueing Model With Two Dedicated Servers.* In Proceedings of the 7th International Conference on Queueing Theory and Network Applications, Kyoto, Japan, August 1-3, 2012.

5. Herwig Bruneel, Willem Mélange, Bart Steyaert, Dieter Claeys and Joris Walraevens. *Impact of Blocking When Customers of Different Classes Are Accommodated in One Common Queue.* In Proceedings of the 1st International Conference on Operations Research and Enterprise Systems, Algarve, Portugal, 2012.

6. Willem Mélange, Joris Walraevens, Dieter Claeys, Bart Steyaert and Herwig Bruneel. *Stability Analysis of Global FCFS and Presorting Service Discipline.* In Proceedings of the 8th International Multi-Conference on Computing in the Global Information Technology, Nice, France, June 21-26, pages 181-187, 2013.

7. Willem Mélange, Joris Walraevens, Dieter Claeys, Bart Steyaert and Herwig Bruneel. *The Impact of Class Clustering on a System with a Global FCFS Service Discipline.* Lecture Notes in Computer Science (Porceedings of the ASMTA 2014 Conference), Budapest, Hungary, June 30 - July 2, 8499:125-139, 2014.

8. Willem Mélange, Joris Walraevens, Dieter Claeys, Bart Steyaert and Herwig Bruneel. *Boundary Problem in a System with Global FCFS and Presorting.* AIP Conference Proceedings, Rhodos, Greece, September 22-28,1648(1), 2014.

9. Willem Mélange, Joris Walraevens, Dieter Claeys, Bart Steyaert and Herwig Bruneel. *Effect of Presorting on the Number of Customers in*

Multiclass Queues with Dedicated Servers and a Global FCFS Service Discipline. In Booklet of Abstracts of the First European Conference on Queueing Theory, Ghent, Belgium, July 10-14, page 33, 2014.

1.6.3 Publications in national conferences

1. Herwig Bruneel, Willem Mélange, Bart Steyaert and Joris Walraevens. *Road splits: smooth or rough passage?*. In Abstracts of the 12th FEA PhD Symposium, Ghent, Belgium, December 7, page 122, 2011.

2. Herwig Bruneel, Willem Mélange, Bart Steyaert and Joris Walraevens. *Analysis of a Queueing Model With Two Dedicated Servers and a Global FCFS Service Discipline*. In Booklet of Abstracts of the 25th Annual Conference of the Belgian Operations Research Society (Orbel25), Ghent, Belgium, February 10-11, pages 95-96, 2011.

3. Herwig Bruneel, Willem Mélange, Dieter Claeys and Joris Walraevens. *A Discrete-Time Queueing Model with Constant Service Times and Blocking*. In Booklet of Abstracts of the 28th Annual Conference of the Belgian Operations Research Society (Orbel28), Mons, Belgium, January 30-31, pages 129-130, 2014.

References

[1] Ward Whitt. *Deciding Which Queue to Join: Some Counterexamples.* Operations Research, 34(1):55–62, 1986.

[2] Nico M. van Dijk. *Why queueing never vanishes.* European Journal of Operations Research, 99(2):463–476, 1997.

[3] Ronald A. Nosek and James P. Wilson. *Queueing Theory and Customer Satisfaction: A Review of Terminology, Trends, and Applications to Pharmacy Practice.* Hospital Pharmacy, 36(3):275–279, 2001.

[4] Richard C. Larson. *Perspectives on Queues: Social Justice and the Psychology of Queueing.* Operations Research, 35(6):895–905, 1987.

[5] Agner K. Erlang. *The Theory of Probabilities and Telephone Conversations.* Nyt Tidsskrift for Matematik, 20(B):33–39, 1909.

[6] John F. Kingman. *The First Erlang Century–and the Next.* Queueing Systems, 63(1-4):3–12, 2009.

[7] Herwig Bruneel. *Nota's bij de lessen Wachtlijntheorie.* University Lecture.

[8] Leonard Kleinrock. *Theory, Volume 1, Queueing Systems.* Wiley-Interscience, 1975.

[9] Robert B. Cooper. *Introduction to Queueing Theory.* Elsevier North Holland, Inc., second edition, 1981.

[10] Kayla C. Lewis. *Forgotten Merits of the Analytic Viewpoint.* Eos, Transactions American Geophysical Union, 94(7):71–72, 2013.

[11] Stewart Robinson. *Simulation: The Practice of Model Development and Use.* Palgrave Macmillan, 2014.

[12] John D. C. Little. *Little's Law as viewed on Its 50th Anniversary.* Operations Research, 59(3):536–549, 2011.

[13] John D. C. Little. *A Proof for the Queueing Formula: $L = \lambda\ W$.* Operations Research, 9(3):383–387, 1961.

[14] Mark W. Harris. *Little's Law: The Science Behind Proper Staffing.* Emergency Physicians Monthly, 2010.

[15] Rajan Suri. *Quick Response Manufacturing: A Companywide Approach to Reducing Lead Times.* Productivity Press, 1998.

[16] Ronald W. Wolff. *Poisson Arrivals See Time Averages.* Operations Research, 30(2):223–231, 1981.

[17] Alois Stutzer and Bruno S. Frey. *Stress that Doesn't Pay: The Commuting Paradox.* Journal of Economics, 110(2):339–366, 2008.

[18] John C. Falcocchio and Herbert S. Levinson. *Road Traffic Congestion: A Concise Guide.* Springer International Publishing, 2015.

[19] Milton D. Harmelink. *Volume Warrants for Left-Turn Lanes at Unsignalized Grade Intersections.* Highway Research Record, (211), 1967.

[20] Joseph C. Oppenlander and Christopher J. Bianchi. *Guidelines for Left-Turn Lanes.* In Institute of Transportation Engineers Meeting, 1990.

[21] Kay Fitzpatrick and Tim Wolff. *Left-Turn Lane Installation Guidelines.* In 2nd Urban Street Symposium: Uptown, Downtown, or Small Town: Designing Urban Streets That Work, 2003.

[22] Ida van Schalkwyk, PE Stover, and G Vergil. *Revisiting Existing Warrants for Left-Turn Lanes at Unsignalized Intersections on Two-Way Roadways.* In Transportation Research Board 86th Annual Meeting, number 07-0784, 2007.

[23] Shinya Kikuchi and Partha Chakroborty. *Analysis of Left-Turn-Lane Warrants at Unsignalized T-Intersections on Two-Lane Roadways.* Transportation Research Record, (1327), 1991.

[24] Partha Chakroborty and Shinya Kikuchi. *Lengths of Left-Turn Lanes at Unsignalized Intersections.* Transportation Research Record, (1500), 1995.

[25] Ponlathep Lertworawanich and Lily Elefteriadou. *Determination of Storage Lengths of Left-Turn Lanes at Unsignalized Intersections Using M/G2/1 Queuing.* Transportation Research Record, (1847), 2003.

[26] G. F. Yeo and B. Weesakul. *Delays to Road Traffic at an Intersection.* Journal of Applied Probability, 1(2):297–310, 1964.

[27] J. C. Tanner. *A Theoretical Analysis of Delays at an Uncontrolled Intersection.* Biometrika, 49(1/2):163–170, 1962.

[28] Ning Wu. *An Approximation for the Distribution of Queue Lengths at Unsignalised Intersections.* In Proceedings of the Second International Symposium on Highway Capacity, volume 2, pages 717–736, 1994.

[29] Dirk Heidemann and Helmut Wegmann. *Queueing at Unsignalized Intersections.* Transportation Research Part B: Methodological, 31(3):239–263, 1997.

[30] BH Cottrell Jr. *The Development of Criteria for the Treatment of Right Turn Movements on Rural Roads.* Technical report, 1981.

[31] Patrick T McCoy, Syed Ataullah, and James A Bonneson. *Guidelines for Right-Turn Lanes on Urban Highways. Final Report.* Technical report, 1993.

2

The impact of the global First-Come-First-Served scheduling

2.1 Introduction

The major aim of this chapter is to quantify the intuitively expected negative impact of the gFCFS service discipline, i.e., all arriving customers are accommodated in one single queue and are served in the order of their arrival regardless of their type, on the performance measures of our system. We want to analyse whether this impact is negligible or if it is important enough to take into account. The results of this chapter give us already a lot of insight into the blocking effect. Moreover, these results can also be considered in a road traffic context as lower and upper bounds for the gain of using of a turn lane (no turn lane and an infinite turn lane). These bounds can already provide us a first insight into the potential gain of a turn lane.

The rest of this chapter can be split into two parts. In section 2.2, we first analyse the system with global FCFS with a focus on the stability of the system, the number of customers in the system and the customer delay. This queueing system is modelled by a continuous-time Markov chain and is analysed using generating functions. Next, section 2.3 is devoted to the comparison of this system with an ideal system, i.e., one without blocking. In this way, we can unveil the impact of the blocking phenomenon.

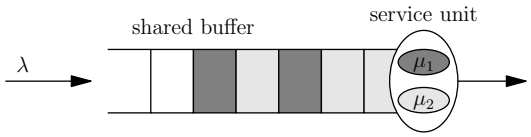


Figure 2.1: Model of the system with global FCFS

2.2 System with a global FCFS service discipline

2.2.1 Mathematical model

We consider a continuous-time queueing model (as shown in Fig. 2.1) with infinite waiting room.

The customers enter the system according to a Poisson arrival process with mean arrival rate λ. We assume customers can be of either of two types, named 1 and 2. The types of consecutive customers are independent, i.e., an arriving customer is of type 1 with probability σ and of type 2 with probability $(1-\sigma)$.

The customers have exponential service times. Server 1 has a service rate of μ_1 and server 2 of μ_2. The servers are dedicated to a given class of customers. Server 1 only serves customers of one type (say type 1) and server 2 serves customers of the other type (type 2).

We assume that customers all queue together and are served in the order of arrival, regardless of the class they belong to. In other words, the service discipline is global FCFS. The global FCFS service discipline creates a blocking effect. When the first two customers in the system are of the same type, the first (oldest) customer in the system of the opposite type is blocked by those customers even though its server is idle. It is this blocking effect we want to emphasize, focus on and analyse in this chapter.

2.2.2 Stability condition

We start this section with introducing the average amount of work (of type 1 and 2) that enters the system per time unit:

$$\rho = \rho_1 + \rho_2 \triangleq \frac{\sigma\lambda}{\mu_1} + \frac{(1-\sigma)\lambda}{\mu_2}.$$

Notice here that our definition is different from the definition of load in most queueing theory literature. Here the load (ρ) is defined as the work arrival rate or traffic intensity. While in literature, load (ρ) is often defined as utilization factor, i.e., the ratio of the time that a system is in use to

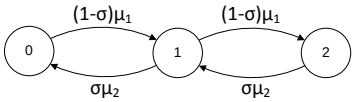


Figure 2.2: Three-state Markov chain to determine the stability condition of the system

the total time that it could be in use. In literature the stability condition is then given by $\rho < 1$. However, using our definition for ρ, the stability condition can then be expressed as

$$\rho < t_0 + 2t_1 + t_2, \tag{2.1}$$

where t_0 represents the fraction of time that only server 1 is working, t_2 the fraction of time that only server 2 is working and t_1 the fraction of time that both servers are working, assuming the system is constantly provided with new customers and, as a result, always at least two customers are present in the system. The system is stable when the average amount of work per time unit that enters the system (ρ) is smaller than the average amount of work the system can serve per time unit, i.e., the average amount of work the system would serve per time unit if there were always at least two customers in the system. When only one server is able to work (whether it be 1 or 2), only one time unit of work per unit time can be served. However when both servers work, two time units of work per unit time are executed, thus explaining (2.1). To determine the fractions of time t_0, t_1 and t_2, we observe that the working servers form a simple three-state Markov chain (Fig. 2.2). State 0 means that only server 1 is working (or, equivalently, the first two customers in our system are of type 1), state 1 that both servers are working (or the first two customers have a different type) and state 2 that only server 2 is working (or, the first two customers in our system are of type 2). This Markov chain is easily solved and t_i is then the fraction of time the Markov chain sojourns in state i ($i = 0, 1, 2$). We find

$$t_0 = \frac{\sigma^2\mu_2^2}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2}, \tag{2.2}$$

$$t_1 = \frac{\sigma(1-\sigma)\mu_1\mu_2}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2}, \tag{2.3}$$

$$t_2 = \frac{(1-\sigma)^2\mu_1^2}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2}. \tag{2.4}$$

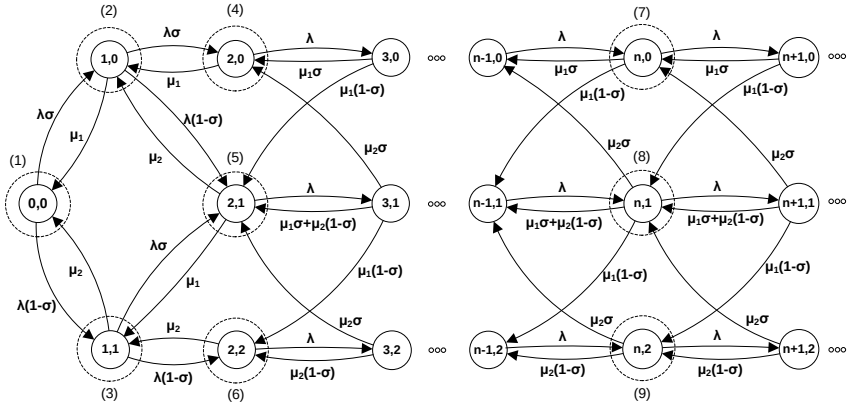


Figure 2.3: State Diagram of the system with global FCFS

Equations (2.1) - (2.4) lead to

$$\rho < \frac{(\sigma\mu_2 + (1-\sigma)\mu_1)^2}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2}. \tag{2.5}$$

or, after some algebra,

$$\lambda < \frac{\left(\frac{\sigma}{\mu_1}\right)^2 - \left(\frac{1-\sigma}{\mu_2}\right)^2}{\left(\frac{\sigma}{\mu_1}\right)^3 - \left(\frac{1-\sigma}{\mu_2}\right)^3}. \tag{2.6}$$

We assume this stability condition to be fulfilled in the remainder of this chapter.

2.2.3 System state diagram and balance equations

The whole system can be described by a continuous-time Markov chain where the state of the system is characterised by the pair (n, m). Here n represents the number of customers in the system (those in service included) and m represents the number of customers of type 2 in the leading customers, i.e., the two customers that are in the system the longest, and thus indicates the types of the two customers at the front (those in service included). In other words, m indicates whether the two customers at the front of the line are both of type 1 ($m = 0$), of alternating types ($m = 1$) or both of type 2 ($m = 2$). Note that in the second case those 2 customers are in service. Otherwise, only the oldest of them is in service and the other is at the front of the queue. The Markov chain is thus a QBD process with three phases (m), and the levels are represented by the number of customers

in the system. Note that only the types of the leading customers are of importance, not the specific order. This is easily explained by the fact that when the leading customers (with at least two customers in the system) are of different types, both servers are able to serve customers regardless of whether these customers are the first or second customer in the system. The types of consecutive customers are also independent so there is no need to keep track of the type of the latest arriving customer in the set of leading customers. By definition, when there are less than two customers in the system ($n = 0, 1$), m represents the number of customers of type 2 in the system. Consequently, states $(0, 1)$, $(0, 2)$ and $(1, 2)$ do not exist. The QBD is drawn in Fig. 2.3.

If we define $p(n, m)$ as the steady-state probability of state (n, m), then we end up with the following balance equations (corresponding to transitions to and from states (1)-(9) in Fig. 2.3):

$$\lambda p(0,0) = \mu_1 p(1,0) + \mu_2 p(1,1) \tag{2.7}$$

$$(\lambda + \mu_1)p(1,0) = \mu_1 p(2,0) + \mu_2 p(2,1) + \sigma\lambda p(0,0) \tag{2.8}$$

$$(\lambda + \mu_2)p(1,1) = \mu_1 p(2,1) + \mu_2 p(2,2) + (1-\sigma)\lambda p(0,0) \tag{2.9}$$

$$(\lambda + \mu_1)p(2,0) = \sigma\left(\mu_1 p(3,0) + \mu_2 p(3,1)\right) + \sigma\lambda p(1,0) \tag{2.10}$$

$$\begin{aligned}(\lambda + \mu_1 + \mu_2)p(2,1) = {} & (1-\sigma)\mu_1 p(3,0) + (\sigma\mu_1 + (1-\sigma)\mu_2)p(3,1) \\ & + \sigma\mu_2 p(3,2) + \lambda((1-\sigma)p(1,0) + \sigma p(1,1))\end{aligned} \tag{2.11}$$

$$\begin{aligned}(\lambda + \mu_2)p(2,2) = {} & (1-\sigma)(\mu_1 p(3,1) + \mu_2 p(3,2)) \\ & + (1-\sigma)\lambda p(1,1)\end{aligned} \tag{2.12}$$

$$\begin{aligned}(\lambda + \mu_1)p(n,0) = {} & \sigma(\mu_1 p(n+1,0) + \mu_2 p(n+1,1)) \\ & + \lambda p(n-1,0), \qquad n \geqslant 3\end{aligned} \tag{2.13}$$

$$\begin{aligned}(\lambda + \mu_1 + \mu_2)p(n,1) = {} & (1-\sigma)\mu_1 p(n+1,0) \\ & + (\sigma\mu_1 + (1-\sigma)\mu_2)p(n+1,1) + \sigma\mu_2 p(n+1,2) \\ & + \lambda p(n-1,1), \qquad n \geqslant 3\end{aligned} \tag{2.14}$$

$$\begin{aligned}(\lambda + \mu_2)p(n,2) = {} & (1-\sigma)(\mu_1 p(n+1,1) + \mu_2 p(n+1,2)) \\ & + \lambda p(n-1,2), \qquad n \geqslant 3\end{aligned} \tag{2.15}$$

For example, the left-hand side of equation (2.13) represents the system leaving state $(n, 0)$ with rate λ (a new customer enters the system) and rate μ_1 (a customer of type 1 leaves the system). Note that a departure of a class 2 customer is impossible when the system is in state $(n, 0)$ since the two leading customers are of type 1. The right-hand side of the equation is a bit more involved. We go to state $(n, 0)$ in three cases. In the first case, this happens with rate λ from state $(n-1, 0)$, i.e., a new customer arrives and finds $n-1$ customers in the system, the two oldest being of type 1. The

arriving customer does not change the leading customers since we assume that there are at least 2 customers in the system when the customer arrives ($n \geqslant 3$). Secondly, the system can go from state $(n+1,0)$ to state $(n,0)$ with rate $\sigma\mu_1$. This happens when a customer of type 1 leaves the system with $n+1$ customers when the leading customers were of type 1 and the leading customers remain of type 1. This is only possible when the new customer in the set of leading customers is of type 1 (with probability σ). The last case is similar to the second case. Here a customer of type 2 leaves when the system is in state $(n+1,1)$, and is replaced by a type 1 customer in the set of leading customers.

2.2.4 Analysis of distributions and moments of the system occupancies

We split this paragraph in two parts. First we tackle the total system occupancy. In a second paragraph, we then investigate the per-type system occupancy. To tackle the analysis of the system occupancies, we make use of pgfs. First, the pgf of the system occupancy is determined. This pgf already gives us straightforwardly some important performance measures (e.g. mean system occupancy). We will also invert the obtained pgfs using partial fraction expansion to obtain (tail asymptotics of) the probability mass function (pmf).

2.2.4.1 Total system occupancy

We start by analysing the total system occupancy, i.e., the number of customers (of all types) in the system. Using the balance equations from Section 2.2.1, the probability generating function $P(z)$ of the total system occupancy can be determined after introducing some partial pgfs for mathematical simplicity. To determine all unknown probabilities, we use the normalization condition and the property that pgfs are bounded inside the closed complex unit disk.

Relation between the pgf $P(z)$ and some partial pgfs

The pgf of the (total) number of customers in the system can be written as

$$P(z) = p(0,0) + z(p(1,0) + p(1,1)) + Q_0(z) + Q_1(z) + Q_2(z), \qquad (2.16)$$

where we introduce the partial pgfs

$$Q_0(z) \triangleq \sum_{n=2}^{\infty} p(n,0)z^n \ , \tag{2.17}$$

$$Q_1(z) \triangleq \sum_{n=2}^{\infty} p(n,1)z^n \ , \tag{2.18}$$

$$Q_2(z) \triangleq \sum_{n=2}^{\infty} p(n,2)z^n. \tag{2.19}$$

Determination of the partial pgfs

Equations (2.13) to (2.15) are multiplied by z^n and summed over all $n \geqslant 3$. We find

$$\begin{aligned}(\lambda+\mu_1)(Q_0(z)-z^2p(2,0)) = \qquad (2.20)\\ \frac{1}{z}\big[\sigma\mu_1(Q_0(z)-z^3p(3,0)-z^2p(2,0))\\ +\sigma\mu_2(Q_1(z)-z^3p(3,1)-z^2p(2,1))\big]+\lambda zQ_0(z)\\ (\lambda+\mu_1+\mu_2)(Q_1(z)-z^2p(2,1)) = \qquad (2.21)\\ \frac{1}{z}\big[(1-\sigma)\mu_1(Q_0(z)-z^3p(3,0)-z^2p(2,0))\\ +(\sigma\mu_1+(1-\sigma)\mu_2)(Q_1(z)-z^3p(3,1)-z^2p(2,1))\\ +\sigma\mu_2(Q_2(z)-z^3p(3,2)-z^2p(2,2))\big]+\lambda zQ_1(z)\\ (\lambda+\mu_2)(Q_2(z)-z^2p(2,2)) = \qquad (2.22)\\ \frac{1}{z}\big[(1-\sigma)\mu_1(Q_1(z)-z^3p(3,1)-z^2p(2,1))\\ +(1-\sigma)\mu_2(Q_2(z)-z^3p(3,2)-z^2p(2,2))\big]+\lambda zQ_2(z).\end{aligned}$$

First, we can eliminate $p(3,m)$ (with $m = 0,1,2$) from (2.20) to (2.22) using (2.10) to (2.12). Then we eliminate $p(2,m)$ with equations (2.8) and (2.9). Notice that in this last step, we have eliminated three unknown probabilities with only two equations. Finally, we use equation (2.7) to eliminate another unknown probability ($p(0,0)$). After eliminating all these

unknown probabilities, this yields

$$[\lambda z^2 - (\lambda + \mu_1)z + \sigma\mu_1]Q_0(z) + \sigma\mu_2 Q_1(z) = \\ - \sigma\lambda p(1,0)z^3 + \sigma[(\lambda + (1-\sigma)\mu_1)p(1,0) - \sigma\mu_2 p(1,1)]z^2, \tag{2.23}$$

$$(1-\sigma)\mu_1 Q_0(z) + [\lambda z^2 - (\lambda + \mu_1 + \mu_2)z + \sigma\mu_1 + (1-\sigma)\mu_2]Q_1(z) \\ + \sigma\mu_2 Q_2(z) = \\ - \lambda((1-\sigma)p(1,0) + \sigma p(1,1))z^3 + [(1-\sigma)((\lambda + (1-\sigma)\mu_1)p(1,0) \\ - \sigma\mu_2 p(1,1)) + \sigma((\lambda + \sigma\mu_2)p(1,1) - (1-\sigma)\mu_1 p(1,0))]z^2, \tag{2.24}$$

$$(1-\sigma)\mu_1 Q_1(z) + [\lambda z^2 - (\lambda + \mu_2)z + (1-\sigma)\mu_2]Q_2(z) = \\ - (1-\sigma)\lambda p(1,1)z^3 + (1-\sigma)[(\lambda + \sigma\mu_2)p(1,1) - (1-\sigma)\mu_1 p(1,0))]z^2. \tag{2.25}$$

What remains is a set of 3 linear equations in the 3 partial pgfs ($Q_m(z)$ with $m = 0, 1, 2$) with only two remaining unknown probabilities ($p(1,0)$ and $p(1,1)$).

Determination of the pgf

We now calculate $P(z)$ from (2.16). Plugging in the solutions of the set of linear equations (2.23) to (2.25), it follows that $P(z)$ is a rational function with a polynomial of degree 6 in the denominator and a polynomial of degree 7 in the numerator. However, calculations show that the coefficient of z^7 in the numerator is zero. It can also be easily seen that $z(z-1)$ is a common factor in the denominator and numerator of $P(z)$. After cancelling the common factors, this leaves us with two polynomials of degree 4 in numerator and denominator, which we will call $N(z)$ and $D(z)$ so that

$$P(z) = \frac{N(z)}{D(z)}. \tag{2.26}$$

These are given by

$$D(z) = \lambda\left[\lambda^3 z^4 - 2\lambda^2(\lambda + \mu_1 + \mu_2)z^3 \right. \\ +\lambda((\lambda + \mu_1 + \mu_2)^2 + 2\lambda((1-\sigma)\mu_2 + \sigma\mu_1) + \mu_1\mu_2)z^2 \\ - (\lambda + \mu_1 + \mu_2)(2\lambda((1-\sigma)\mu_2 + \sigma\mu_1) + \mu_1\mu_2)z \\ \left. +\lambda(\lambda + \mu_1 + \mu_2)^2 + \mu_1\mu_2((\lambda + \mu_1 + \mu_2) - \lambda\sigma(1-\sigma))\right], \tag{2.27}$$

$$N(z) = f(1-\sigma, \mu_2, \mu_1, z)p(1,0) + f(\sigma, \mu_1, \mu_2, z)p(1,1), \tag{2.28}$$

with

$$\begin{aligned}
f(w,x,y,z) &= f_0(w,x,y) + f_1(w,x,y)z + f_2(w,x,y)z^2 + f_3(w,x,y)z^3 \\
&\quad + f_4(w,x,y)z^4, \\
f_4(w,x,y) &= -\lambda^3 xw, \\
f_3(w,x,y) &= 2\lambda^3 xw + \lambda^2\left(-y\,(x+y)\,w\,(1-w) + wx^2 - (1-w)\,y^2\right), \\
f_2(w,x,y) &= \lambda\left(-xw\lambda^2 + \left(w\,(1-w)\,x^2 + 2xy + \left(1-w^2\right)y^2\right)(\lambda+y)\right. \\
&\quad \left. + \left(-2x\left((1-w)\,y + wx\right) + \left(1-w^2\right)y^2\right)\lambda\right), \\
f_1(w,x,y) &= \left((x-y)^2\,w^2 - xyw\,(1-w) - y^2\right)\lambda^2 \\
&\quad + y\left(-\,(x+2y)\,(x-y)\,w - 2y\,(x+y)\right)\lambda - xy^2\,(x+y), \\
f_0(w,x,y) &= y\left(\left((x-y)^2\,w^2 + xyw\,(1+w) + (1-2w)\,y^2\right)\lambda\right. \\
&\quad \left. + y\left(wx^2 + (1-w)\,xy\right)\right).
\end{aligned}$$

The two remaining unknown probabilities ($p(1,0)$ and $p(1,1)$) can be determined, in general, by invoking the well-known property that pgfs such as $P(z)$ are bounded inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z-plane, at least when the stability condition (2.5) of the queueing system is met (only in such a case our analysis was justified and $P(z)$ can be viewed as a legitimate pgf). It is clear that the zeroes inside the closed unit disk of the denominator should also be zeroes of the numerator of (2.26), as $P(z)$ must remain bounded in those points. First we calculate all zeroes of the denominator. By means of the method of Ferrari [1], the four zeroes are of the form

$$\frac{a_1 \pm_s \sqrt{a_2 \pm_t 2\sqrt{a_3}}}{2\lambda} \tag{2.29}$$

with

$$\begin{aligned}
a_1 &= \lambda + \mu_1 + \mu_2, \\
a_2 &= (\lambda + \mu_1 + \mu_2)^2 - 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2) - 2\mu_1\mu_2, \\
a_3 &= \mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2),
\end{aligned}$$

where the signs $\pm_s$ or $\pm_t$ can be plus or minus (so four options for four zeroes). We can prove that the only zero of these four zeroes that is inside the closed unit disk when the stability condition is met, named $\hat{z}_0$ in the rest of the dissertation, is the zero for $\pm_s = -$ and $\pm_t = +$ (see Appendix A). For ease of notation in the rest of this dissertation we will call the zero where $\pm_s = -$ and $\pm_t = -$, $\hat{z}_1$, for $\pm_s = +$ and $\pm_t = -$, $\hat{z}_2$ and for $\pm_s = +$ and $\pm_t = +$, $\hat{z}_3$. We can prove that all zeroes are on the positive real axis (see Appendix A).

The requirement that the numerator should vanish at $\hat{z}_0$ yields a linear equation for the two remaining unknowns. A second linear equation can be obtained by invoking the normalization condition of the pgf $P(z)$, i.e., the condition $P(1) = 1$. In general, the two unknown probabilities can be found as the solutions of the two established linear equations. Substitution of the obtained values in (2.26) then leads to a fully determined and explicit expression for the steady-state pgf $P(z)$ of the system occupancy which is of the following (partial fractions) form

$$P(z) = b_0 + \frac{b_1}{z - \hat{z}_1} + \frac{b_2}{z - \hat{z}_2} + \frac{b_3}{z - \hat{z}_3}, \tag{2.30}$$

with

$$b_j = \frac{N(\hat{z}_j)}{D'(\hat{z}_j)} \text{ with } j = 1, 2, 3, \tag{2.31}$$

$$b_0 = 1 + \sum_{j=1}^{3} \frac{b_j}{\hat{z}_j - 1}. \tag{2.32}$$

where the zero $\hat{z}_0$ inside the unit circle is cancelled out.

Determination of the pmf

From these results the pmf $p(n)$ can be determined. The pmf is then fully and exactly given by

$$p(n) = \begin{cases} b_0 - b_1 \cdot \hat{z}_1^{-1} - b_2 \cdot \hat{z}_2^{-1} - b_3 \cdot \hat{z}_3^{-1}, & \text{n} = 0 \\ -b_1 \cdot \hat{z}_1^{-n-1} - b_2 \cdot \hat{z}_2^{-n-1} - b_3 \cdot \hat{z}_3^{-n-1}, & n > 0 \end{cases}. \tag{2.33}$$

The pmf is thus a mixture of geometric distributions and extra mass in zero.

A similar approach can be adopted for the partial pgfs. The corresponding partial pmfs are given by

$$p(n, m) = -b_1(m) \cdot \hat{z}_1^{-n-1} - b_2(m) \cdot \hat{z}_2^{-n-1} - b_3(m) \cdot \hat{z}_3^{-n-1}, \quad n > 1 \tag{2.34}$$

with

$$b_j(m) = \frac{N_m(\hat{z}_j)}{D'(\hat{z}_j)} \text{ with } j = 1, 2, 3 \text{ and } m = 0, 1, 2. \tag{2.35}$$

where $N_m(z)$ is the numerator of $Q_m(z)$ and $D(z)$ is the denominator of $Q_m(z)$.

2.2.4.2 Per-type system occupancies

We are not only interested in the distribution of the total system occupancy but also in the distributions for both customer classes separately. The joint pgf $P(z_1, z_2)$ of the number of customers (of type 1 and 2) in the system is given by

$$P(z_1,z_2) = p(0,0) + z_1 p(1,0) + z_2 p(1,1) + z_1^2 \frac{Q_0(\sigma z_1 + (1-\sigma)z_2)}{(\sigma z_1 + (1-\sigma)z_2)^2} + z_1 z_2 \frac{Q_1(\sigma z_1 + (1-\sigma)z_2)}{(\sigma z_1 + (1-\sigma)z_2)^2} + z_2^2 \frac{Q_2(\sigma z_1 + (1-\sigma)z_2)}{(\sigma z_1 + (1-\sigma)z_2)^2}, \quad (2.36)$$

where the last three terms can be explained by noticing that we know the types of the first two customers in our system and that all other customers are of type 1 with probability σ or of type 2 with probability $1-\sigma$. The pgfs of the system occupancies of both customer types separately are then given by

$$P_1(z_1) = P(z_1, 1), \quad (2.37)$$

$$P_2(z_2) = P(1, z_2) \quad (2.38)$$

and again the pmfs of the system occupancies of both customer types separately can be derived and are given by

$$p_1(n_1) = \sum_{j=1}^{3} \sum_{m=0}^{2} -\frac{(\hat{z}_j - (1-\sigma))^{2-m} b_j(m)}{\sigma^{3-m} \hat{z}_j^2} \left(\frac{\hat{z}_j - (1-\sigma)}{\sigma} \right)^{-n_1-1}, \quad (2.39)$$

$$p_2(n_2) = \sum_{j=1}^{3} \sum_{m=0}^{2} -\frac{(\hat{z}_j - \sigma)^{2-m} b_j(m)}{(1-\sigma)^{3-m} \hat{z}_j^2} \left(\frac{\hat{z}_j - \sigma}{1-\sigma} \right)^{-n_2-1}, \quad (2.40)$$

where $p_i(n_i)$ is the probability that there are n_i customers of type i in the system with $n_i \geq 0$.

2.2.5 Analysis of the distribution and moments of the system delays of a customer

2.2.5.1 System delay of a random customer

To tackle the analysis of the delay of a random customer, Laplace transforms are used. First the LST of the delay of a random customer is determined. Then this Laplace transform is inverted to find the probability density function (pdf).

Relation between the delay of a customer and conditional delay of a customer

Define $s_{n,m}(t)$ as the pdf of the system delay (S) of a customer given that the customer sees the state (n,m) upon arrival. Using the PASTA property (see 1.3.3.6) and (2.34) we get the pdf $s(t)$ of the total system delay

$$\begin{aligned} s(t) &= \sum_{n,m} p(n,m) s_{n,m}(t) \qquad (2.41)\\ &= p(0,0)s_{0,0}(t) + p(1,0)s_{1,0}(t) + p(1,1)s_{1,1}(t) \\ &\quad + \sum_{m=0}^{2}\sum_{n=2}^{\infty} s_{n,m}(t)\left(-b_1(m)\hat{z}_1^{-n-1} - b_2(m)\hat{z}_2^{-n-1} - b_3(m)\hat{z}_3^{-n-1}\right). \end{aligned}$$

To determine this pdf, we first compute its Laplace transform, namely

$$s^*(\theta) = p(0,0)s^*_{0,0}(\theta) + p(1,0)s^*_{1,0}(\theta) + p(1,1)s^*_{1,1}(\theta) + \sum_{j=1}^{3} H_j(\theta), \quad (2.42)$$

where $s^*_{n,m}(\theta)$ are the Laplace transforms of the above defined $s_{n,m}(t)$,

$$H_j(\theta) = \sum_{m=0}^{2} -b_j(m) G_{j,m}(\theta) \qquad (2.43)$$

and

$$G_{j,m}(\theta) = \sum_{n=2}^{\infty} s^*_{n,m}(\theta)\hat{z}_j^{-n-1}. \qquad (2.44)$$

Determination of conditional delays of a customer

To construct the Laplace transform $s^*(\theta)$, some reasoning in the Laplace domain yields the following recursive relations for $n > 2$

$$\begin{aligned} s^*_{n,0}(\theta) &= \frac{\mu_1}{\mu_1+\theta}\left(\sigma s^*_{n-1,0}(\theta) + (1-\sigma)s^*_{n-1,1}(\theta)\right), \qquad (2.45)\\ s^*_{n,1}(\theta) &= \frac{\mu_1}{\mu_1+\mu_2+\theta}\left(\sigma s^*_{n-1,1}(\theta) + (1-\sigma)s^*_{n-1,2}(\theta)\right) \\ &\quad + \frac{\mu_2}{\mu_1+\mu_2+\theta}\left(\sigma s^*_{n-1,0}(\theta) + (1-\sigma)s^*_{n-1,1}(\theta)\right), \qquad (2.46)\\ s^*_{n,2}(\theta) &= \frac{\mu_2}{\mu_2+\theta}\left(\sigma s^*_{n-1,1}(\theta) + (1-\sigma)s^*_{n-1,2}(\theta)\right). \qquad (2.47) \end{aligned}$$

Equation (2.45) can be understood as follows: the delay of a customer that arrives when the system is in state $(n,0)$ equals the sum of an exponentially distributed service time with rate μ_1 and the delay of a (virtual) customer

arriving in a state with one less customer, i.e., state $(n-1,m)$, where $m=0$ with probability σ and $m=1$ with probability $1-\sigma$. A similar reasoning leads to equations (2.46) and (2.47). We get by multiplying (2.45), (2.46) and (2.47) by $\hat{z}_j^{-n-1}$ and summing over all $n \geqslant 2$

$$\begin{aligned}
(\mu_1+\theta)(G_{j,0}(\theta)-\hat{z}_j^{-3}s^*_{2,0}(\theta)) &= \sigma\mu_1\hat{z}_j^{-1}G_{j,0}(\theta) \qquad (2.48)\\
&\quad +(1-\sigma)\mu_1\hat{z}_j^{-1}G_{j,1}(\theta),\\
(\mu_1+\mu_2+\theta)(G_{j,1}(\theta)-\hat{z}_j^{-3}s^*_{2,1}(\theta)) &= \sigma\mu_1\hat{z}_j^{-1}G_{j,1}(\theta)+\sigma\mu_2\hat{z}_j^{-1}G_{j,0}(\theta)\\
&\quad +(1-\sigma)\mu_1\hat{z}_j^{-1}G_{j,2}(\theta)\\
&\quad +(1-\sigma)\mu_2\hat{z}_j^{-1}G_{j,1}(\theta), \qquad (2.49)\\
(\mu_2+\theta)(G_{j,2}(\theta)-\hat{z}_j^{-3}s^*_{2,2}(\theta)) &= \sigma\mu_2\hat{z}_j^{-1}G_{j,1}(\theta) \qquad (2.50)\\
&\quad +(1-\sigma)\mu_2\hat{z}_j^{-1}G_{j,2}(\theta).
\end{aligned}$$

Multiplying (2.48) by $-b_j(0)$, (2.49) by $-b_j(1)$, (2.50) by $-b_j(2)$ and adding them yields

$$\begin{aligned}
&\theta H_j(\theta)+(\mu_1+\theta)b_j(0)\hat{z}_j^{-3}s^*_{2,0}(\theta)+(\mu_1+\mu_2+\theta)b_j(1)\hat{z}_j^{-3}s^*_{2,1}(\theta)\\
&+(\mu_2+\theta)b_j(2)\hat{z}_j^{-3}s^*_{2,2}(\theta)=\\
&\qquad G_{j,0}(\theta)\left[\mu_1 b_j(0)-\sigma\mu_1 b_j(0)\hat{z}_j^{-1}-\sigma\mu_2 b_j(1)\hat{z}_j^{-1}\right]\\
&\quad +G_{j,1}(\theta)\left[(\mu_1+\mu_2)b_j(1)-(1-\sigma)\mu_1 b_j(0)\hat{z}_j^{-1}-\sigma\mu_1 b_j(1)\hat{z}_j^{-1}\right.\\
&\quad \left.-(1-\sigma)\mu_2 b_j(2)\hat{z}_j^{-1}-\sigma\mu_2 b_j(2)\hat{z}_j^{-1}\right]\\
&\quad +G_{j,2}(\theta)\left[\mu_2 b_j(2)-(1-\sigma)\mu_1 b_j(1)\hat{z}_j^{-1}-(1-\sigma)\mu_2 b_j(2)\hat{z}_j^{-1}\right]. \qquad (2.51)
\end{aligned}$$

To further simplify (2.51), we first insert (2.34) in (2.13) - (2.15) and divide by $\hat{z}_j^{-n-1}$. We get

$$\begin{aligned}
(\lambda+\mu_1)\left(-b_j(0)\right) &= \sigma(\mu_1\left(-b_j(0)\hat{z}_j^{-1}\right)+\mu_2\left(-b_j(1)\hat{z}_j^{-1}\right)\\
&\quad +\lambda\left(-b_j(0)\hat{z}_j\right), \qquad (2.52)\\
(\lambda+\mu_1+\mu_2)\left(-b_j(1)\right) &= (1-\sigma)\mu_1\left(-b_j(0)\hat{z}_j^{-1}\right)+\sigma\mu_2\left(-b_j(2)\hat{z}_j^{-1}\right)\\
&\quad +(\sigma\mu_1+(1-\sigma)\mu_2)\left(-b_j(1)\hat{z}_j^{-1}\right)\\
&\quad +\lambda\left(-b_j(1)\hat{z}_j\right), \qquad (2.53)\\
(\lambda+\mu_2)\left(-b_j(2)\right) &= (1-\sigma)(\mu_1\left(-b_j(1)\hat{z}_j^{-1}\right)+\mu_2\left(-b_j(2)\hat{z}_j^{-1}\right))\\
&\quad +\lambda\left(-b_j(2)\hat{z}_j\right), \qquad (2.54)
\end{aligned}$$

or rewritten

$$-\lambda b_j(0)\,(1-\hat{z}_j) = \mu_1 b_j(0) - \sigma\mu_1 b_j(0)\hat{z}_j^{-1} - \sigma\mu_2 b_j(1)\hat{z}_j^{-1}, \tag{2.55}$$

$$\begin{aligned}-\lambda b_j(1)\,(1-\hat{z}_j) = {} & (\mu_1+\mu_2)b_j(1) - (1-\sigma)\mu_1 b_j(0)\hat{z}_j^{-1} - \sigma\mu_1 b_j(1)\hat{z}_j^{-1} \\ & - (1-\sigma)\mu_2 b_j(1)\hat{z}_j^{-1} - \sigma\mu_2 b_j(2)\hat{z}_j^{-1},\end{aligned} \tag{2.56}$$

$$-\lambda b_j(2)\,(1-\hat{z}_j) = \mu_2 b_j(2) - (1-\sigma)\mu_1 b_j(1)\hat{z}_j^{-1} - (1-\sigma)\mu_2 b_j(2)\hat{z}_j^{-1}. \tag{2.57}$$

Now by using (2.55)-(2.57), we can simplify the right-hand side of (2.51) to get

$$\begin{aligned}&\theta H_j(\theta) + (\mu_1+\theta)b_j(0)\hat{z}_j^{-3}s^*_{2,0}(\theta) + (\mu_1+\mu_2+\theta)b_j(1)\hat{z}_j^{-3}s^*_{2,1}(\theta) \\ &+ (\mu_2+\theta)b_j(2)\hat{z}_j^{-3}s^*_{2,2}(\theta) = \\ &\quad - G_{j,0}(\theta)\lambda b_j(0)(1-\hat{z}_j) - G_{j,1}(\theta)\lambda b_j(1)(1-\hat{z}_j) - G_{j,2}(\theta)\lambda b_j(2)(1-\hat{z}_j).\end{aligned} \tag{2.58}$$

Using (2.43) finally yields

$$\begin{aligned}H_j(\theta) = \frac{-1}{\theta-\lambda(1-\hat{z}_j)}\big[&(\mu_1+\theta)b_j(0)\hat{z}_j^{-3}s^*_{2,0}(\theta) + (\mu_2+\theta)b_j(2)\hat{z}_j^{-3}s^*_{2,2}(\theta) \\ &+(\mu_1+\mu_2+\theta)b_j(1)\hat{z}_j^{-3}s^*_{2,1}(\theta)\big].\end{aligned} \tag{2.59}$$

At this point, we are able to express $s^*(\theta)$ in (2.42) in terms of the boundary Laplace transforms ($s^*_{0,0}(\theta)$,$s^*_{1,0}(\theta)$,$\cdots$, $s^*_{2,2}(\theta)$) using (2.59). These boundary Laplace transforms are given by

$$s^*_{0,0}(\theta) = \frac{\sigma\mu_1}{\mu_1+\theta} + \frac{(1-\sigma)\mu_2}{\mu_2+\theta}, \tag{2.60}$$

$$s^*_{1,0}(\theta) = \sigma\left(\frac{\mu_1}{\mu_1+\theta}\right)^2 + (1-\sigma)\frac{\mu_2}{\mu_2+\theta}, \tag{2.61}$$

$$s^*_{1,1}(\theta) = \sigma\frac{\mu_1}{\mu_1+\theta} + (1-\sigma)\left(\frac{\mu_2}{\mu_2+\theta}\right)^2, \tag{2.62}$$

$$s^*_{2,0}(\theta) = \sigma\left(\frac{\mu_1}{\mu_1+\theta}\right)^3 + (1-\sigma)\frac{\mu_1\mu_2}{(\mu_1+\theta)(\mu_2+\theta)}, \tag{2.63}$$

$$\begin{aligned}s^*_{2,1}(\theta) = \frac{1}{\mu_1+\mu_2+\theta}\Bigg(&\sigma\left(\mu_2\left(\frac{\mu_1}{\mu_1+\theta}\right)^2 + \mu_1\frac{\mu_1}{\mu_1+\theta}\right) \\ &+(1-\sigma)\left(\mu_2\frac{\mu_2}{\mu_2+\theta} + \mu_1\left(\frac{\mu_2}{\mu_2+\theta}\right)^2\right)\Bigg),\end{aligned} \tag{2.64}$$

$$s^*_{2,2}(\theta) = \sigma\frac{\mu_1\mu_2}{(\mu_1+\theta)(\mu_2+\theta)} + (1-\sigma)\left(\frac{\mu_2}{\mu_2+\theta}\right)^3. \tag{2.65}$$

Note that the tagged arriving customer is able to start its service when it is the second customer in the system if the first customer is of a different type. Otherwise, the customer has to wait until the (older) customer of the same type is served. Equation (2.63), for instance, can be understood as follows: by definition, the tagged customer sees state (2,0) on arrival. Thus there are two customers in the system of type 1. If the tagged customer is also of type 1 (with probability σ) then the two customers that are already in the system, have to be served first. All three customers have an exponentially distributed service time with rate μ_1. Contrary, if the tagged customer is of type 2 (with probability $1-\sigma$), then the tagged customer can start its service as soon as one customer has left the system. The first customer in the system (of type 1) has an exponentially distributed service time with rate μ_1 and the tagged customer has an exponentially distributed service time with rate μ_2 in that case. The other equations can be understood in a similar way.

Determination of the density of the delay

After inserting (2.59)-(2.65) into (2.42) and after some simplifications, we get

$$\begin{aligned} s^*(\theta) = \sum_{j=1}^{3} \frac{1}{\theta + \lambda(\hat{z}_j - 1)} \sum_{m=0}^{2} \Bigg[& \frac{-b_j(m)}{\hat{z}_j^3 \left(\mu_1 - \lambda(\hat{z}_j - 1)\right)^2 \left(\mu_2 - \lambda(\hat{z}_j - 1)\right)^2} \\ & \cdot \left(\frac{m}{2}(3-m)\mu_1 + \frac{-m^2+m+2}{2}\mu_2 - \lambda(\hat{z}_j - 1) \right) \\ & \cdot \Bigg((1-\sigma)\mu_1^{\frac{m^2-3m+2}{2}} \mu_2^{m+1} \left(\mu_1 - \lambda(\hat{z}_j - 1)\right)^{\frac{-m^2+m+4}{2}} \\ & + \sigma \mu_1^{3-m} \mu_2^{\frac{m}{2}(m-1)} \left(\mu_2 - \lambda(\hat{z}_j - 1)\right)^{\frac{-m^2+3m+2}{2}} \Bigg) \Bigg] \end{aligned} \tag{2.66}$$

From the Laplace transform of the probability density function (pdf), we can again derive some performance measures of practical importance. For instance, the mean system delay can be found as $T = -\left.\frac{ds^*(\theta)}{d\theta}\right|_{\theta=0}$. The pdf can be derived easily by taking the inverse Laplace transform and is given

by

$$s(t) = \sum_{j=1}^{3} e^{-\lambda(\hat{z}_j-1)t} \sum_{m=0}^{2} \left[\frac{-b_j(m)}{\hat{z}_j^3 \left(\mu_1 - \lambda(\hat{z}_j-1)\right)^2 \left(\mu_2 - \lambda(\hat{z}_j-1)\right)^2} \right.$$
$$\cdot \left(\frac{m}{2}(3-m)\mu_1 + \frac{-m^2+m+2}{2}\mu_2 - \lambda(\hat{z}_j-1) \right)$$
$$\cdot \left((1-\sigma)\mu_1^{\frac{m^2-3m+2}{2}} \mu_2^{m+1} \left(\mu_1 - \lambda(\hat{z}_j-1)\right)^{\frac{-m^2+m+4}{2}} \right.$$
$$\left. \left. +\sigma\mu_1^{3-m}\mu_2^{\frac{m}{2}(m-1)} \left(\mu_2 - \lambda(\hat{z}_j-1)\right)^{\frac{-m^2+3m+2}{2}} \right) \right]. \tag{2.67}$$

Hence, the probability density function $s(t)$ is a hyperexponential distribution.

2.2.5.2 Per-type customer delays

We are not only interested in the pdf of the global system delay but also in the pdfs for both customer types separately. Define $s_{1,n,m}(t)$ as the conditional probability density function (pdf) of the system delay (S_1) of a customer of type 1 given that the customer sees the state (n,m) upon arrival. Using the PASTA property and (2.34), we get for the probability density function $s_1(t)$ of the system delay

$$\begin{aligned} s_1(t) &= \sum_{n,m} p(n,m) s_{1,n,m}(t) \\ &= p(0,0)s_{1,0,0}(t) + p(1,0)s_{1,1,0}(t) + p(1,1)s_{1,1,1}(t) \\ &\quad + \sum_{m=0}^{2}\sum_{n=2}^{\infty} s_{1,n,m}(t) \left(-b_1(m)\hat{z}_1^{-n-1} - b_2(m)\hat{z}_2^{-n-1} - b_3(m)\hat{z}_3^{-n-1}\right). \end{aligned} \tag{2.68}$$

Notice here that $s_{1,n,m}(t)$ has the same iterative equations as $s_{n,m}(t)$ for $n > 2$ because as long as the tagged customer is not among the first two customers of the system, the type of the tagged customer has no influence on the system behaviour. A similar analysis can thus be followed as in section 2.2.5.1 to determine the Laplace transform of the system delay of a customer of type 1, leading to

$$s_1^*(\theta) = p(0,0)s_{1,0,0}^*(\theta) + p(1,0)s_{1,1,0}^*(\theta) + p(1,1)s_{1,1,1}^*(\theta) + \sum_{j=1}^{3} H_j(\theta). \tag{2.69}$$

To fully determine $s_1^*(\theta)$, we need the following boundary values

$$s_{1,0,0}^*(\theta) = \frac{\mu_1}{\mu_1+\theta}, \tag{2.70}$$

$$s_{1,1,0}^*(\theta) = \left(\frac{\mu_1}{\mu_1+\theta}\right)^2, \tag{2.71}$$

$$s_{1,1,1}^*(\theta) = \frac{\mu_1}{\mu_1+\theta}, \tag{2.72}$$

$$s_{1,2,0}^*(\theta) = \left(\frac{\mu_1}{\mu_1+\theta}\right)^3, \tag{2.73}$$

$$s_{1,2,1}^*(\theta) = \frac{1}{\mu_1+\mu_2+\theta}\left(\mu_2\left(\frac{\mu_1}{\mu_1+\theta}\right)^2 + \mu_1\frac{\mu_1}{\mu_1+\theta}\right), \tag{2.74}$$

$$s_{1,2,2}^*(\theta) = \frac{\mu_1\mu_2}{(\mu_1+\theta)(\mu_2+\theta)} \tag{2.75}$$

which are again easily deduced. The pdf can be derived easily by taking the inverse Laplace transform and is given by

$$\begin{aligned} s_1(t) = \sum_{j=1}^{3} e^{-\lambda(\hat{z}_j-1)t} \sum_{m=0}^{2} \Bigg[& \frac{-b_j(m)}{\hat{z}_j^3\left(\mu_1-\lambda(\hat{z}_j-1)\right)^2\left(\mu_2-\lambda(\hat{z}_j-1)\right)^2} \\ & \cdot \left(\frac{m}{2}(3-m)\mu_1 + \frac{-m^2+m+2}{2}\mu_2 - \lambda(\hat{z}_j-1)\right) \\ & \cdot \left(\mu_1^{3-m}\mu_2^{\frac{m}{2}(m-1)}\left(\mu_2-\lambda(\hat{z}_j-1)\right)^{\frac{-m^2+3m+2}{2}}\right)\Bigg]. \end{aligned} \tag{2.76}$$

Thus, the pdf $s_1(t)$ is again a hyperexponential distribution.

Using the symmetry of our system, the pdf of the system delay of customers of type 2 can be found by swapping μ_1 with μ_2, σ with $1-\sigma$ and m with $2-m$ in (2.76). The pdf is thus given by

$$\begin{aligned} s_2(t) = \sum_{j=1}^{3} e^{-\lambda(\hat{z}_j-1)t} \sum_{m=0}^{2} \Bigg[& \frac{-b_j(m)}{\hat{z}_j^3\left(\mu_1-\lambda(\hat{z}_j-1)\right)^2\left(\mu_2-\lambda(\hat{z}_j-1)\right)^2} \\ & \cdot \left(\frac{m}{2}(3-m)\mu_1 + \frac{-m^2+m+2}{2}\mu_2 - \lambda(\hat{z}_j-1)\right) \\ & \cdot \left(\mu_1^{\frac{m^2-3m+2}{2}}\mu_2^{m+1}\left(\mu_1-\lambda(\hat{z}_j-1)\right)^{\frac{-m^2+m+4}{2}}\right)\Bigg]. \end{aligned} \tag{2.77}$$

2.3 Comparison of models and numerical examples

In this section we want to quantify the impact of the gFCFS service discipline. We do this by comparing it to a system where there is no blocking.

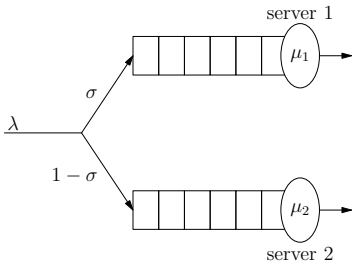


Figure 2.4: Network model of the system without global FCFS

We first write down the (straightforward) analysis of the latter system.

2.3.1 Ideal reference system without blocking

When the global FCFS restriction is dropped and replaced by the service discipline FCFS for each of the types separately (i.e., a customer can be served if no customers of its own type are in front of it), the system is equivalent to a network of two parallel, separate queues (see Fig. 2.4) where each customer upon arrival goes immediately to the queue (waiting room) in front of its own server. Since this is a Jackson network, this network can be split in two simple $M \mid M \mid 1$-queues and the marginal system content distributions of those queues can be multiplied to get the joint distribution. The distribution of a simple $M \mid M \mid 1$-queue is well-known from many books about queueing theory, e.g. [2].

The stability condition is given by

$$\lambda < min(\frac{\mu_1}{\sigma}, \frac{\mu_2}{1-\sigma}). \tag{2.78}$$

The pmfs of the system occupancies are given by

$$p_1(n_1) = \left(1 - \frac{\sigma\lambda}{\mu_1}\right)\left(\frac{\sigma\lambda}{\mu_1}\right)^{n_1}, \tag{2.79}$$

$$p_2(n_2) = \left(1 - \frac{(1-\sigma)\lambda}{\mu_2}\right)\left(\frac{(1-\sigma)\lambda}{\mu_2}\right)^{n_2}, \tag{2.80}$$

$$\begin{aligned} p(n) &= \sum_{i=0}^{n} p_1(i)p_2(n-i) \qquad (2.81)\\ &= \left(1 - \frac{\sigma\lambda}{\mu_1}\right)\left(1 - \frac{(1-\sigma)\lambda}{\mu_2}\right)\frac{\left(\frac{\sigma\lambda}{\mu_1}\right)^{n+1} - \left(\frac{(1-\sigma)\lambda}{\mu_2}\right)^{n+1}}{\frac{\sigma\lambda}{\mu_1} - \frac{(1-\sigma)\lambda}{\mu_2}}, \end{aligned}$$

where $p_i(n_i)$ is the probability that there are n_i customers of type i in the system with $n_i \geq 0$ and $p(n)$ is the probability that there are n customers in the system with $n \geq 0$. The pdfs of the customer delay are given by

$$\begin{aligned} s_1(t_1) &= \left(1 - \frac{\sigma\lambda}{\mu_1}\right) \mu_1 e^{-\left(1-\frac{\sigma\lambda}{\mu_1}\right)\mu_1 t_1}, && (2.82) \\ s_2(t_2) &= \left(1 - \frac{(1-\sigma)\lambda}{\mu_2}\right) \mu_2 e^{-\left(1-\frac{(1-\sigma)\lambda}{\mu_2}\right)\mu_2 t_2}, && (2.83) \\ s(t) &= \sigma s_1(t) + (1-\sigma) s_2(t) \\ &= \sigma \left(1 - \frac{\sigma\lambda}{\mu_1}\right) \mu_1 e^{-\left(1-\frac{\sigma\lambda}{\mu_1}\right)\mu_1 t} \\ &\quad + (1-\sigma) \left(1 - \frac{(1-\sigma)\lambda}{\mu_2}\right) \mu_2 e^{-\left(1-\frac{(1-\sigma)\lambda}{\mu_2}\right)\mu_2 t}. && (2.84) \end{aligned}$$

where $s_i(t_i)$ is the pdf of the customers of type i with $t_i \geq 0$ and $s(t)$ is the pdf of a random customer with $t \geq 0$.

2.3.2 Numerical comparison

In the remainder of this section, we discuss the results obtained in the previous sections, from a quantitative and a qualitative perspective, by means of some numerical examples. Before discussing the results, we introduce a new parameter

$$\omega \triangleq \frac{\frac{\sigma}{\mu_1}}{\frac{\sigma}{\mu_1} + \frac{1-\sigma}{\mu_2}} = \frac{\rho_1}{\rho_1 + \rho_2}. \tag{2.85}$$

This parameter will allow us to interpret the results more intuitively; it represents the relative load of customers of type 1. This parameter can, for instance, be introduced in the stability conditions (2.5) and (2.78), yielding

$$\rho < \frac{1}{1 - \omega(1-\omega)} \tag{2.86}$$

and

$$\rho < min\left(\frac{1}{\omega}, \frac{1}{1-\omega}\right). \tag{2.87}$$

In the remainder, we will first show the impact of the global FCFS service discipline on the total system. Secondly, we will zoom in on the impact of the balance in the system over the two types. Finally, we will demonstrate that global FCFS can have a major impact on the customers that form the minority.

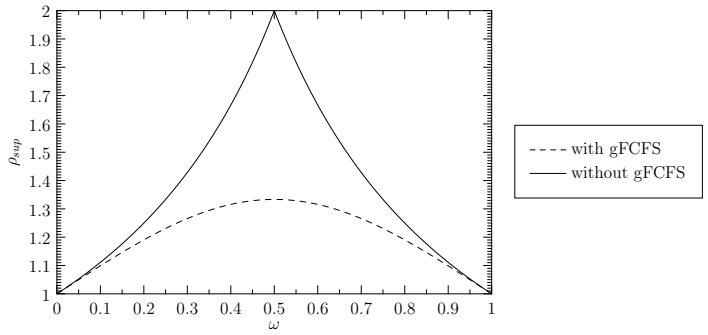


Figure 2.5: ρ_{sup}, least upper bound of the set of values ρ where the system is stable versus parameter ω

Impact of global FCFS

Fig. 2.5 shows ρ_{sup}, least upper bound of the set of values ρ where the system is stable versus parameter ω, with global FCFS as service discipline (gFCFS) and without global FCFS. It is clear that the system with global FCFS always performs "worse" than the system without this restriction especially for values of ω around $\omega = \frac{1}{2}$. An observation that is also confirmed in Fig. 2.9 where the mean system occupancy versus parameter ρ with $\sigma = \frac{30}{31}$, $\mu_1 = 20$ and $\mu_2 = 2$ (and thus $\omega = \frac{3}{4}$) is shown and in Fig. 2.10 where the mean delay versus parameter ω with $\rho = 1$, $\mu_1 = 2$ and $\mu_2 = 2$ is shown.

Impact of the load balance between customers of type 1 and customers of type 2 (parameter ω)

In Fig. 2.5, we notice that the achievable total throughput is maximum in both systems (with and without gFCFS) when the load is equally balanced. We also see that the effect of gFCFS is devastating in this case. When the relative loads are very unbalanced, say $\omega < 0.2$ and $\omega > 0.8$, the difference between the systems with and without gFCFS becomes negligible if we look at the maximum achievable total throughput. This is as expected since both systems approach the single-server case if the majority of load is of the same class.

Fig. 2.6 presents the mean delay versus parameter ω with $\mu_1 = 20$ and $\mu_2 = 2$ for different values of ρ. Opposite to what we stated above and in the system without blocking, we see that a well balanced system ($\omega = 0.5$) gives no longer the best result when we deal with small total loads (ρ). A system where the fastest server gets a higher relative load performs better than the well balanced system. When the total load increases, the best performing system becomes more balanced. This is again intuitively clear,

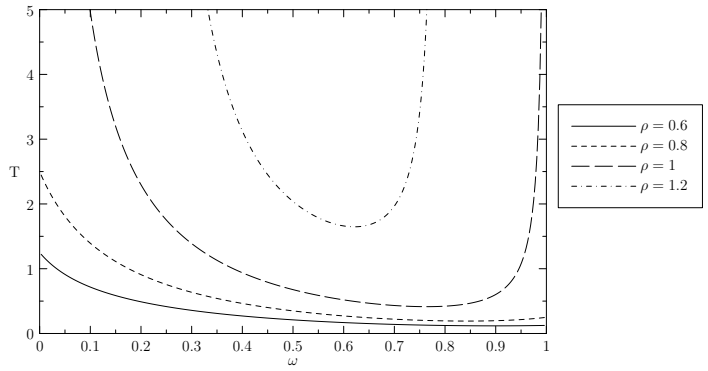


Figure 2.6: Mean delay versus parameter ω with $\mu_1 = 20$ and $\mu_2 = 2$

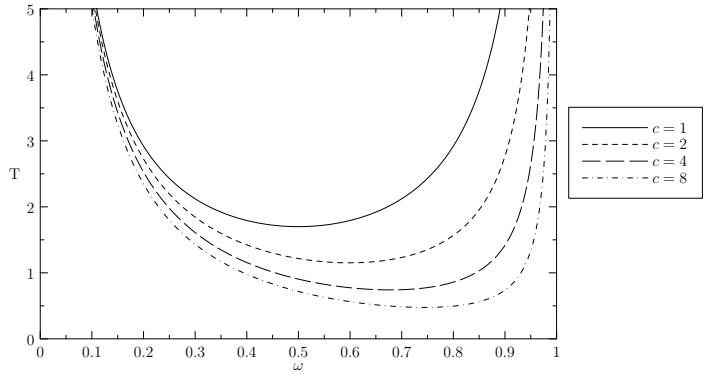


Figure 2.7: Mean delay versus parameter ω with $\rho = 1$, $\mu_1 = c\mu_2$ and $\mu_2 = 2$

since in cases that the demand of the arrival stream is considerably less than what can be handled by 1 server, the question of whether the second server is also active or not, is not very relevant. The blocking effect dominates the effect of both servers working. The fastest server should therefore get some preference (if possible). When the demand increases, the question of whether the second server is also active or not, becomes more relevant and thus a well balanced system becomes more preferable. We also see that, for high loads, the system is very sensitive to ω, while this is less so for lightly loaded systems.

Fig. 2.7 shows the mean delay versus parameter ω with $\rho = 1$, $\mu_1 = c\mu_2$ and $\mu_2 = 2$. We see similar effects as in the previous case. The faster one of the servers is (relative to the speed of the other server), the more it should be preferred.

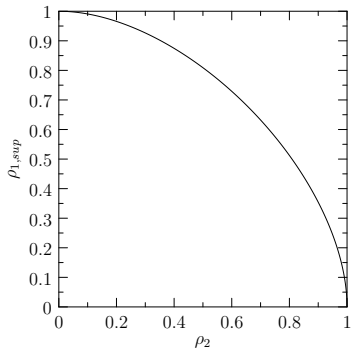


Figure 2.8: $\rho_{1,sup}$, least upper bound of the set of values ρ_1 where the system is stable versus parameter ρ_2

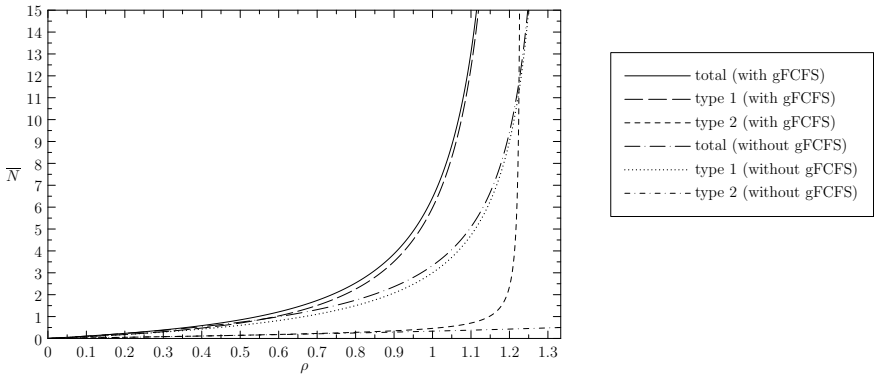


Figure 2.9: Mean system occupancy versus parameter ρ with $\sigma = \frac{30}{31}$, $\mu_1 = 20$ and $\mu_2 = 2$ ($\omega = \frac{3}{4}$)

Impact of customers of one type on customer performance of the other type

Fig. 2.8 represents $\rho_{1,sup}$, the least upper bound of the set of values ρ_1 where the system is stable versus parameter ρ_2 in a system with gFCFS. Here we can already see the major difference between systems with and without gFCFS. In the system without gFCFS, the maximum allowable loads of the two types of customers are independent. On the other hand, in the system with gFCFS, there exist a relationship between both types of customers. For example, in the system with gFCFS, as we can see in Fig. 2.8, when the load of customers of type 2 is 0.8, then the load of customers of type 1 is maximum 0.51 for the system to remain stable. Whereas in the system without gFCFS, the load of customers of type 1 can still be as much as 1.

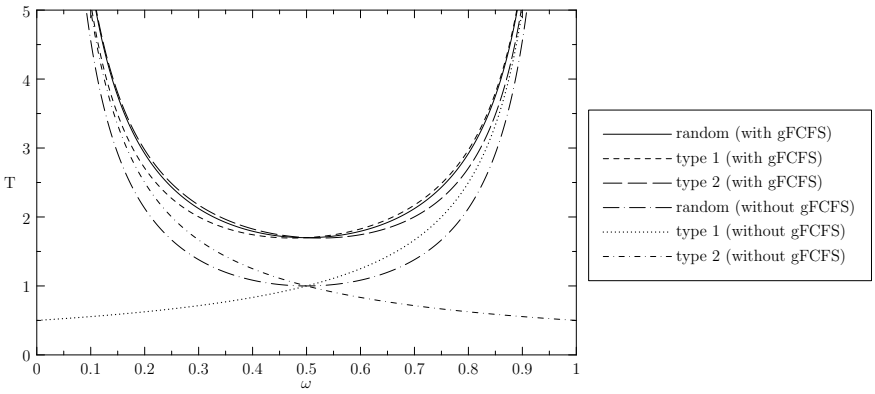


Figure 2.10: Mean delay versus parameter ω with $\rho = 1$, $\mu_1 = 2$ and $\mu_2 = 2$

From Fig. 2.9 we notice again that the system without gFCFS performs better than the system with gFCFS for ρ approaching 1. However when ρ is small the difference is negligible. This is intuitively clear since, for small ρ, the demand of the arrival stream is considerably less than the traffic that can be handled by 1 server, and therefore, whether the second server is active or not, is not crucial. The system with gFCFS also has a negative influence on customers of type 2. Again we can see that there exists a relationship between both types of customers. Customers of type 2 get stuck in the same queue as customers of type 1 and therefore suffer for high loads. This is even more clear in Fig. 2.10. As stated above, when the load is out of balance, the difference between the systems with and without gFCFS becomes negligible if we look at the mean system time of a random customer. However we see that the system with gFCFS has a big negative impact on the customers that form the minority. While in the system without gFCFS these customers have a mean delay that is very small, they now have a system time that approaches that of the customers that form the majority. We can say that there is some kind of levelling of the system times of both types of customers.

Use for dimensioning purposes

An advantage of sharing a buffer is to save space. However, if this means that we need to enforce a global FCFS service discipline, this no longer holds. Fig. 2.11 shows the tail probability of the system occupancy with $\rho = 1$, $\sigma = \frac{30}{31}$, $\mu_1 = 20$ and $\mu_2 = 2$ (and thus $\omega = \frac{3}{4}$). The tail probability is the probability that the number of customers in the system is larger than a value i. The tail probability can be considered as an approximate value for

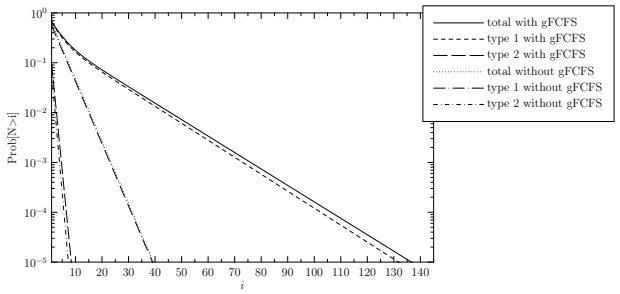


Figure 2.11: The tail probability of the system occupancy with $\rho = 1$, $\sigma = \frac{30}{31}$, $\mu_1 = 20$ *and* $\mu_2 = 2$ *(and thus* $\omega = \frac{3}{4}$*)*

the loss probability in a system with finite storage capacity equal to i places, and which can be used for dimensioning purposes. For example, the required buffer size is 39 and 8 (together 47) for a loss ratio of 10^{-5} when assuming two separate buffers (with FCFS service discipline). If you can share the buffer and do not have to restrict to a global FCFS scheduling, you need a buffer size of 40. However, the required buffer size is 137 for a loss ratio of 10^{-5} when assuming one buffer with a global FCFS service discipline. So it is beneficial to share a buffer when there is no restriction to a global FCFS service discipline (40 versus 47 required buffer size). This is no longer the case when there is a restriction to a global FCFS service discipline (47 versus 137 required buffer size). Notice that in the example, the difference is so extreme because of the large difference in load of customers of type 1 and load of customers of type 2.

In a traffic context hard and soft constraints can be suggested for optimization. A hard constraint is one that must be satisfied at all times. A soft constraint is a want to be satisfied as much as possible if the cost for doing so is not too great [3]. The warrants suggested in Section 1.4 are hard constraints. A turn lane is warranted when the hard constraint is not fulfilled. Then the junction is considered too unsafe (prone to accidents) and other costs are not considered. However, a lot of other cost functions could be considered. A lot of research is being done about the economic and environmental cost of congestion. These costs should be compared to the real cost for constructing the turn lane using soft constraints. However, the determination of cost functions goes beyond the scope of this dissertation. In the rest of this dissertation, we will only consider hard constraints.

A lot of the results in this chapter can be used to construct soft or hard constraints to determine whether or not a turn lane is warranted (and to already estimate the potential impact). Here, we will consider only two possibles constraints, but other constraints could be constructed using the

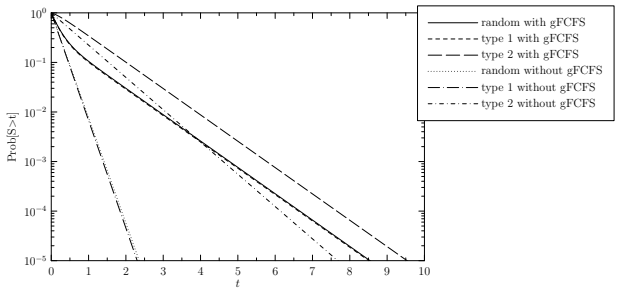


Figure 2.12: The tail probability of the system delay with $\rho = 1$, $\sigma = \frac{30}{31}$, $\mu_1 = 20$ *and* $\mu_2 = 2$ *(and thus* $\omega = \frac{3}{4}$*)*

results in this chapter.

The tail probability of the systems occupancy can be used for such a hard constraint and thus to warrant a turn lane. One possibility is to determine the probability that the traffic jam caused by the blocking effect spreads to other junctions causing a domino effect. This probability should be less than a certain well-chosen threshold value. In the example in Fig. 2.11 when choosing a threshold value of 10^{-5} when there is a road for 50 vehicles, a turn lane is warranted. We can also see that this threshold is feasible since in the most ideal case (two lanes), the threshold value is not exceeded.

Another possibility is to determine a well-chosen threshold value for the probability that the delay of a random vehicle is more than a certain value. For example, the probability that a vehicle's delay is more than 5 is less than 10^{-5}. In Fig. 2.12, where the tail probability of the delay of a customer with $\rho = 1$, $\sigma = \frac{30}{31}$, $\mu_1 = 20$ and $\mu_2 = 2$ (and thus $\omega = \frac{3}{4}$) is shown, a turn lane is warranted. Again, we can see that this threshold is feasible. Also notice here that the majority in number (or customers of type 1) has the most influence on the delay of a random customer. However, their delay is the most influenced by the majority of load (or customers of type 2).

References

[1] Jean-Pierre Tignol. *Galois' Theory of Algebraic Equations.* 2001.

[2] Herwig Bruneel. *Nota's bij de lessen Wachtlijntheorie.* University Lecture.

[3] JW Kendall. *Hard and Soft Constraints in Linear Programming.* Omega, 3(6):709–715, 1975.

3 The impact of class clustering

3.1 Introduction

In this chapter, we shift focus to the effect of class clustering, i.e., the way customers of any given type have a tendency to "arrive back-to-back". Class clustering is a concept that often is neglected in literature to keep the model as simple as possible, but in this chapter we want to demonstrate that it is not always possible to treat this concept negligently. It is already intuitively clear that when the customers arrive with alternating types, less blocking will occur than when types alternate only very rarely. We quantify this effect in this chapter.

We already want to point out here that this chapter fits into the part of the research question of studying the model thoroughly and broaden our comprehension about the model. In a traffic context, arrivals are considered to be random and no class clustering exist (see Section 1.4). Consequently, this chapter adds little in a traffic context other than broadening our general comprehension of the model. However, other applications of our models exist where class clustering can be of importance. For example, at a security checkpoint (e.g., at an international airport or train station) people are usually body-searched by someone of the same gender. As a result, when a group of friends of the same gender arrive, the people of the opposite gender behind them may have to wait until the whole group has been checked, even when the other security person is available, at least when it is not allowed to overtake at the security checkpoint (which is often the case for security

reasons).

The rest of this chapter can be split into two parts. In section 3.2, we first analyse the most simple model capturing the concept of class clustering. This is a model with only one cluster parameter. We do this with a focus on the stability of the system, the number of customers in the system and the customer delay. This system is modelled by a continuous-time Markov chain and is solved using a compensation approach. Next, in section 3.3, we extent this system with an extra cluster parameter (two cluster parameters or one for each type). The model is analysed with a focus on the stability of the system and the number of customers in the system. This system is modelled by a continuous-time Markov chain and is solved using generating functions.

3.2 One cluster parameter

In this section, we keep the model that incorporates the concept of class clustering as simple as possible by exploiting the symmetry in the system and by only introducing one extra parameter.

3.2.1 Mathematical model

We consider a continuous-time queueing model with infinite waiting room. There are two types (classes) of customers which are being served at rate μ (exponential service times) independent of the type. Each of the two servers is dedicated to a given class of customers. In this case, server 1 always serves customers of type 1 and server 2 always serves customers of type 2. The customers are served in their order of arrival, regardless of the class they belong to (gFCFS).

The customers enter the system according to a Poisson arrival process with mean arrival rate λ. In this chapter, the major aim is to estimate the impact of the degree of class clustering in the arrival process on this two-class two-server system. To explicitly model this, we assume a first-order Markovian type of correlation between the types of two consecutively arriving customers, which basically means that the probability that the next customer belongs to a given class depends on the class of the previous customer. We denote by α the probability that the next customer has *the same type as the previous one*, and by $1 - \alpha$ the probability that the next customer belongs to *the opposite type as the previous one*. The parameter α can then be considered as a measure of the degree of class clustering in the arrival process, and will therefore be referred to as the "cluster parameter" in the sequel. It is easily seen that the size of a cluster of customers of

the same type, i.e., the number of consecutive customers of any given type between two customers of the opposite type, is geometrically distributed with parameter α and mean value $1/(1-\alpha)$. From a conceptual point of view, the only price we pay with this choice is that we can only study cases where both classes of customers are equiprobable and thus both types of customers account for half of the total load of the system.

3.2.2 Stability condition

We start this section with introducing the average amount of work that enters the system per time unit:

$$\rho \triangleq \frac{\lambda}{\mu}.$$

The stability condition can then be expressed as

$$\rho < t_0 + 2t_1, \tag{3.1}$$

where t_0 represents the fraction of time when one server is working and t_1 is the fraction of time when both servers are working when the system is constantly fed with new customers. Indeed, the system is stable when the average amount of work per time unit that enters the system (ρ) is smaller than the average amount of work the system can execute per time unit, i.e., the average amount of work the system would execute per time unit when it would be constantly provided with new customers. When only one server is able to work, only one time unit work per time unit can be executed. However when both servers can work, two time units work per time unit can be executed, thus explaining (3.1). We can also rewrite (3.1) as

$$\rho < 1 + t_1, \tag{3.2}$$

which we can interpret as follows: there will always be at least one server working when the system is constantly provided with new customers and only a fraction of time a second server is working. To determine the fractions of time t_0 and t_1, we conceive that the number of working servers forms a simple two-state Markov chain. The rate to go from the state where only one server (state 0) is working to the state where both servers are working (state 1), is $(1-\alpha)\mu$; namely a rate μ to end the service in state 0 multiplied with the probability $1-\alpha$ that the next two first customers of our system are of opposite types. The rate to go from state 1 to state 0 equals μ; namely a rate 2μ to end the service in state 1 multiplied with the probability $\frac{1}{2}$ that the next customer to be served is of the opposite type of the departed customer. This last probability is the sum of (i) the probability $\frac{1}{2}\alpha$ that

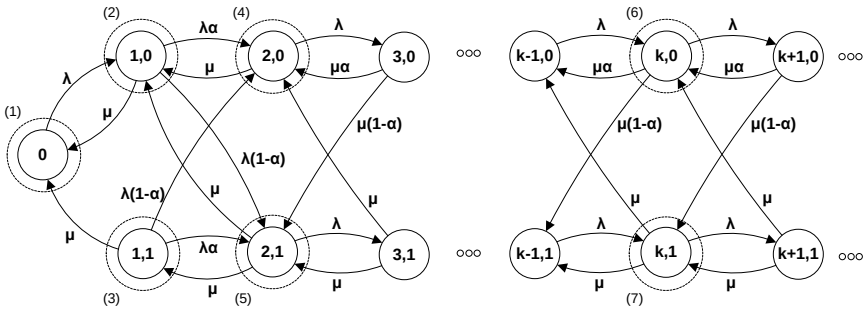


Figure 3.1: The state diagram

the oldest customer is served first (with probability $\frac{1}{2}$ since both type of customers have the same exponential service time) and the next customer to be served is of the opposite type of the departed customer (same type as the second oldest customer), and (ii) the probability $\frac{1}{2}(1-\alpha)$ that the second oldest customer is served first and the next customer to be served is of the opposite type of the departed customer (different type of the second oldest customer). The time t_0 is then the fraction of time the Markov chain sojourns in state 0 and is given by

$$t_0 = \frac{1}{2-\alpha}. \tag{3.3}$$

Similarly, the time t_1 is the fraction of time the Markov chain sojourns in state 1 and is given by

$$t_1 = \frac{1-\alpha}{2-\alpha}. \tag{3.4}$$

Equations (3.1) and (3.2) lead to

$$\rho < \frac{3-2\alpha}{2-\alpha}. \tag{3.5}$$

We assume this stability condition to be fulfilled in the remainder of the section.

3.2.3 System state diagram and balance equations

This system can be described by a continuous-time Markov process. We describe the state of the system by the pair (n, m) where n represents the number of customers in the system and m indicates whether the leading customers, i.e., the two customers that are in the system the longest (those in service included), are of the same type ($m = 0$) or not ($m = 1$). For example, the state $(n, 0)$ means that the types of the leading customers are

of the same type and a total of n customers resides in the system. This is thus a QBD process with two phases (m) and where the levels represent the number of customers in the system. Note, we are only interested in the (difference of) types of the leading customers. We stress that the exact type of the leading customers is of no importance, only whether they are different. The first reason for this is that the cluster parameter α denotes the probability that the next customer is of the same type as the previous customer, regardless of the type of this customer. Secondly, the service times of both types of customers have the same distribution, namely an exponential one with mean $\frac{1}{\mu}$. When we have only one customer in the system ($n = 1$) we interpret the state differently. When the system is in the state $(1, 0)$, the customer left in the system has the same type as the last customer that arrived in the system. Similarly, $(1, 1)$ means that the customer left in the system has a different type than the last customer that arrived in the system. Indeed, the last customer may have already left the system because customers are able to overtake each other in the service units (by having a shorter service time). Notice also that it is not necessary to split state (0) in a $(0, 0)$ and a $(0, 1)$ state because we always enter state $(1, 0)$ from both states. Thus, due to our choice for modelling the system as a symmetric one (the arrival process is only determined by the total arrival rate λ and the cluster parameter α and the service times are equally distributed as well), the state diagram of Fig. 3.1 emerges.

Define $p(n, m)$ as the steady-state probability to be in state (n, m). Then from Fig. 3.1 (see transitions to and from states (6) and (7)), we obtain the following balance equations for $p(n, m)$, for the repeating portion of our Markov chain ($n > 2$),

$$(\lambda + \mu)p(n, 0) = \alpha\mu p(n + 1, 0) + \mu p(n + 1, 1) + \lambda p(n - 1, 0), \tag{3.6}$$

$$(\lambda + 2\mu)p(n, 1) = (1 - \alpha)\mu p(n + 1, 0) + \mu p(n + 1, 1) + \lambda p(n - 1, 1). \tag{3.7}$$

The following boundary equations can be obtained similarly (observe transitions to and from states $(1) - (5)$ in Fig. 3.1, resp.)

$$\lambda p(0) = \mu p(1, 0) + \mu p(1, 1), \tag{3.8}$$

$$(\lambda + \mu)p(1, 0) = \mu p(2, 0) + \mu p(2, 1) + \lambda p(0), \tag{3.9}$$

$$(\lambda + \mu)p(1, 1) = \mu p(2, 1), \tag{3.10}$$

$$(\lambda + \mu)p(2, 0) = \alpha\mu p(3, 0) + \mu p(3, 1) + \alpha\lambda p(1, 0) + (1 - \alpha)\lambda p(1, 1), \tag{3.11}$$

$$(\lambda + 2\mu)p(2, 1) = (1 - \alpha)\mu p(3, 0) + \mu p(3, 1) + (1 - \alpha)\lambda p(1, 0) + \alpha\lambda p(1, 1). \tag{3.12}$$

with $p(0)$ the steady-state probability to be in state (0).

3.2.4 Analysis of the distribution and moments of the system occupancy

QBD processes have a well known geometric relation [1] and the pmf of the system occupancy is of the form

$$p(n,m) = \sum_{j=0}^{1} C_j y_j(m) x_j^n, \quad n > 1 \text{ and } m = 0, 1. \tag{3.13}$$

The approach to solve this problem is inspired by ideas from [2–5]. We start by searching for basic solutions (with $n > 2$) of the balance equations (3.6) and (3.7), assuming the form

$$p(n,m) = y(m)x^n. \tag{3.14}$$

The general solution (3.13) can be expressed as a linear combination of these basic solutions. Using the boundary conditions and normalization condition, all remaining unknowns can be determined.

Determination of the basic solutions

Substituting (3.14) in (3.6) and (3.7), and dividing by x^{n-1} yields

$$(\alpha\mu x^2 - (\lambda+\mu)x + \lambda)y(0) + \mu x^2 y(1) = 0, \tag{3.15}$$

$$(1-\alpha)\mu x^2 y(0) + (\mu x^2 - (\lambda + 2\mu)x + \lambda)y(1) = 0. \tag{3.16}$$

Equations (3.15) and (3.16) form a linear homogeneous system of equations in $y(0)$ and $y(1)$ for a given x. Since the system is homogeneous, x must be chosen such that the determinant of the system is zero. A direct determination of the values of x with this property is often computationally unattractive. Therefore, we will transform (3.15) and (3.16) into a single differential equation using the generating function

$$Y(z) = y(0) + y(1)z, \tag{3.17}$$

after which we solve this single differential equation using separation of variables. To do this we start by multiplying (3.15) by z^0, (3.16) by z^1 (to retain all information) and we sum both, yielding

$$\begin{aligned}(1-\alpha)\mu x^2 z(Y(z) - zY'(z)) + (\mu\alpha x^2 - (\lambda+\mu)x + \lambda)Y(z) \\ + ((1-\alpha)\mu x^2 - \mu x))zY'(z) + \mu x^2 Y'(z) = 0.\end{aligned} \tag{3.18}$$

After some rewriting we get

$$\begin{aligned}\frac{Y'(z)}{Y(z)} &= \frac{(1-\alpha)\mu x^2 z + \mu\alpha x^2 - (\lambda+\mu)x + \lambda}{(1-\alpha)\mu x^2 z^2 - \mu x((1-\alpha)x - 1)z - \mu x^2} \\ &= \frac{A(x)}{z - z_1(x)} + \frac{1 - A(x)}{z - z_2(x)}.\end{aligned} \tag{3.19}$$

where

$$z_1(x) = \frac{(1-\alpha)x - 1 + \sqrt{(1-\alpha)(5-\alpha)x^2 - 2(1-\alpha)x + 1}}{(1-\alpha)x},$$
$$z_2(x) = \frac{(1-\alpha)x - 1 - \sqrt{(1-\alpha)(5-\alpha)x^2 - 2(1-\alpha)x + 1}}{(1-\alpha)x},$$

and

$$A(x) = \frac{1}{2}\left(\frac{-\mu(1+\alpha)x^2 + (2\lambda+3\mu)x - 2\lambda}{\mu x\sqrt{(1-\alpha)(5-\alpha)x^2 - 2(1-\alpha)x + 1}} + 1\right).$$

The general solution of (3.19) is given by

$$Y(z) = K(z - z_1(x))^{A(x)}(z - z_2(x))^{1-A(x)},$$

with K an arbitrary constant. The goal of introducing $Y(z)$ was to find the x wherefore the determinant of the system given in (3.15) and (3.16) is zero, so that the system has a nontrivial solution, i.e., a nonzero solution. This is equivalent to imposing the condition that the generating function $Y(z)$ is a nonzero polynomial. In our case, the exponents $A(x)$ and $1 - A(x)$ should both be non-negative integers. This condition is only met when $A(x) = j$ $(j = 0, 1)$ because assigning another integer value to j would cause one of the exponents to become negative. So now we get two equations (one equation for each j) and, provided the system is stable, each of these two equations yields exactly one root (x_j) in the interval $(0, 1)$. This is easily proven by looking at $A(x)$ and its first derivative $A'(x)$. With some straightforward algebra we can prove that $A(x)$ is a continuous function that first increases from $-\infty$ and then decreases to 1. This makes that there is exactly one root between $(0, 1)$ for both equations. After determination of these x_j, we can determine the $y_j(m)$ corresponding with each given x_j, i.e., determine the nontrivial solution of the homogeneous system given in (3.15) and (3.16). To find a particular solution we use the initial values $y_j(1) = 1$.

Determination general solution and performances measures

Having found basic solutions (3.14) of the balance equations (3.6) and (3.7), we can express the general solution as a linear combination of these basic solutions (for $n > 1$ and $m = 0, 1$),

$$p(n, m) = \sum_{j=0}^{1} C_j y_j(m) x_j^n. \tag{3.20}$$

To have a fully specified distribution we still need to specify five unknowns (C_0, C_1, $p(0)$, $p(1, 0)$ and $p(1, 1)$). To determine these unknowns, we use

the boundary equations (3.8) to (3.12) and the normalizing condition,

$$p(0) + p(1,0) + p(1,1) + \sum_{n=2}^{\infty} \sum_{m=0}^{1} p(n,m) = 1. \tag{3.21}$$

Notice that we can drop one of the boundary equations since the boundary equations are dependent (the normalization condition replaces this boundary equation).

Knowing the probability mass function (pmf), we can derive some performance measures of practical importance using (3.8)-(3.12) and (3.20). For example, the mean system occupancy can be found as

$$\begin{aligned} \overline{N} &= p(1,0) + p(1,1) + \sum_{n=2}^{\infty} \sum_{m=0}^{1} np(n,m) \\ &= \sum_{j=0}^{1} \sum_{m=0}^{1} \frac{C_j y_j(m) x_j^2((m+1)\mu(1-x_j)^2 + \lambda(2-x_j))}{(1-x_j)^2\lambda}. \end{aligned} \tag{3.22}$$

Using Little's Law, the mean system delay equals

$$\begin{aligned} T &= \frac{\overline{N}}{\lambda} \\ &= \sum_{j=0}^{1} \sum_{m=0}^{1} \frac{C_j y_j(m) x_j^2((m+1)\mu(1-x_j)^2 + \lambda(2-x_j))}{(1-x_j)^2\lambda^2}. \end{aligned} \tag{3.23}$$

Finally the tail probability is given by

$$\text{Prob}[N > i] = \begin{cases} \sum_{j=0}^{1} \sum_{m=0}^{1} \frac{C_j y_j(m) x_j^2((m+1)\mu(1-x_j)+\lambda)}{(1-x_j)\lambda}, & i = 0 \\ \sum_{j=0}^{1} \sum_{m=0}^{1} C_j y_j(m) \frac{x_j^{i+1}}{1-x_j}, & i > 0. \end{cases} \tag{3.24}$$

where N is the random variable for the system occupancy.

3.2.5 Analysis of the distribution and moments of the system delay of a random customer

To tackle the analysis of the delay of a random customer, Laplace transforms are used. First the LST of the delay of a random customer is determined. Then this Laplace transform is inverted to find the pdf.

Relation between the delay of a customer and conditional delay of a customer

Define $s_{n,m}(t)$ as the probability density function of the system delay of a customer given that the customer sees the state (n,m) on arrival. Using

the PASTA property and (3.20) we get for the probability density function of the system delay

$$\begin{aligned} s(t) = \frac{Prob[t < S < t + dt]}{dt} &= \sum_{n,m} p(n,m) s_{n,m}(t) \\ &= p(0)s_0(t) + p(1,0)s_{1,0}(t) + p(1,1)s_{1,1}(t) \\ &\quad + \sum_{j=0}^{1} \sum_{m=0}^{1} C_j y_j(m) \sum_{n=2}^{\infty} s_{n,m}(t) x_j^n. \end{aligned} \tag{3.25}$$

To determine this pdf, we first compute its Laplace transform, namely

$$s^*(\theta) = p(0)s_0^*(\theta) + p(1,0)s_{1,0}^*(\theta) + p(1,1)s_{1,1}^*(\theta) + \sum_{j=0}^{1} C_j H_j(\theta), \tag{3.26}$$

where $s_{n,m}^*(\theta)$ are the Laplace transforms of the above defined $s_{n,m}(t)$,

$$H_j(\theta) = \sum_{m=0}^{1} y_j(m) G_{j,m}(\theta), \tag{3.27}$$

and

$$G_{j,m}(\theta) = \sum_{n=2}^{\infty} s_{n,m}^*(\theta) x_j^n. \tag{3.28}$$

Determination of the conditional delays of a customer

To construct the Laplace transform $s^*(\theta)$, some reasoning in the Laplace domain yields the following recursive relations

$$s_{n,0}^*(\theta) = \frac{\mu}{\mu + \theta} \left(\alpha s_{n-1,0}^*(\theta) + (1-\alpha) s_{n-1,1}^*(\theta) \right), \tag{3.29}$$

$$s_{n,1}^*(\theta) = \frac{2\mu}{2\mu + \theta} \left(\frac{1}{2} s_{n-1,0}^*(\theta) + \frac{1}{2} s_{n-1,1}^*(\theta) \right). \tag{3.30}$$

Equation (3.29) can be understood as follows: the delay of a customer arriving when the system is in state $(n,0)$ equals the sum of an exponentially distributed service time with rate μ and the delay of a (virtual) customer arriving in a state with one less customer, i.e., state $(n-1,m)$, where $m = 0$ with probability α and $m = 1$ with probability $1 - \alpha$. A similar reasoning leads to equation (3.30). We get by multiplying (3.29) and (3.30) with x_j^n and summing for all $n > 2$

$$(\mu + \theta)(G_{j,0}(\theta) - x_j^2 s_{2,0}^*) = \alpha\mu x_j G_{j,0}(\theta) + (1-\alpha)\mu x_j G_{j,1}(\theta), \tag{3.31}$$

$$(2\mu + \theta)(G_{j,1}(\theta) - x_j^2 s_{2,1}^*) = \mu x_j G_{j,0}(\theta) + \mu x_j G_{j,1}(\theta), \tag{3.32}$$

Multiplying (3.31) with $y_j(0)$, (3.32) with $y_j(1)$ and adding both yields

$$\begin{aligned}\theta H_j(\theta) - (\mu+\theta)x_j^2 y_j(0)s^*_{2,0} - (2\mu+\theta)x_j^2 y_j(1)s^*_{2,1} = \\ G_{j,0}(\theta)((\alpha\mu x_j - \mu)y_j(0) + \mu x_j y_j(1)) \\ + G_{j,1}(\theta)((1-\alpha)\mu x_j y_j(0) + (\mu x_j - 2\mu)y_j(1)).\end{aligned} \tag{3.33}$$

After using (3.15) and (3.16) we get

$$\begin{aligned}\theta H_j(\theta) - (\mu+\theta)x_j^2 y_j(0)s^*_{2,0} - (2\mu+\theta)x_j^2 y_j(1)s^*_{2,1} = \\ G_{j,0}(\theta)\lambda(1-\frac{1}{x_j})y_j(0) + G_{j,1}(\theta)\lambda(1-\frac{1}{x_j})y_j(1).\end{aligned} \tag{3.34}$$

Using (3.28) finally yields

$$H_j(\theta) = \frac{1}{\theta - \lambda(1-\frac{1}{x_j})}\sum_{m=0}^{1}((m+1)\mu+\theta)y_j(m)x_j^2 s^*_{2,m}(\theta). \tag{3.35}$$

A this point, we are able to express s^* in (3.26) in terms of the boundary Laplace transforms $(s^*_0(\theta), s^*_{1,0}(\theta), \cdots, s^*_{2,1}(\theta))$ using (3.35). These boundary Laplace transforms are given by

$$s^*_0(\theta) = \frac{\mu}{\mu+\theta}, \tag{3.36}$$

$$s^*_{1,0}(\theta) = (1-\alpha)\frac{\mu}{\mu+\theta} + \alpha(\frac{\mu}{\mu+\theta})^2, \tag{3.37}$$

$$s^*_{1,1}(\theta) = \alpha\frac{\mu}{\mu+\theta} + (1-\alpha)(\frac{\mu}{\mu+\theta})^2, \tag{3.38}$$

$$s^*_{2,0}(\theta) = \alpha(\frac{\mu}{\mu+\theta})^3 + (1-\alpha)(\frac{\mu}{\mu+\theta})^2, \tag{3.39}$$

$$s^*_{2,1}(\theta) = \frac{1}{2}\frac{2\mu}{2\mu+\theta}\frac{\mu}{\mu+\theta} + \frac{1}{2}\frac{2\mu}{2\mu+\theta}(\frac{\mu}{\mu+\theta})^2, \tag{3.40}$$

which are easily deduced. Note that the tagged customer is able to start his service when he is the second customer in the system if the first customer is of a different type. Otherwise, the customer has to wait until he is the first customer in the system. After inserting (3.36)-(3.40) into (3.26) and after some simplifications, we get

$$s^*(\theta) = \sum_{j=0}^{1}\sum_{m=0}^{1} c_j(m)\frac{1}{\theta - \lambda\left(1-\frac{1}{x_j}\right)} \tag{3.41}$$

with

$$c_j(m) = \frac{C_j\mu^2 x_j^2 y_j(m)\left((m+1)\mu + (1-\alpha+m\alpha)\lambda\left(1-\frac{1}{x_j}\right)\right)}{\left(\mu - \lambda\left(1-\frac{1}{x_j}\right)\right)^2} \tag{3.42}$$

and by taking the inverse Laplace transform, we find

$$s(t) = \sum_{j=0}^{1} \sum_{m=0}^{1} c_j(m) e^{\lambda\left(1-\frac{1}{x_j}\right)t}. \tag{3.43}$$

Thus, the probability density function $s(t)$ is a hyperexponential distribution.

Determination of performances measures

Knowing the probability density function (pdf), we can again derive some performance measures of practical importance. For example, the mean system delay can be found as

$$\begin{aligned} T &= \int_0^\infty t s(t) dt \\ &= \sum_{j=0}^{1} \sum_{m=0}^{1} c_j(m) \frac{1}{\left(\lambda\left(1-\frac{1}{x_j}\right)\right)^2}. \end{aligned} \tag{3.44}$$

While (3.44) looks a bit different than (3.23), it can be proven that both are identical by inserting the explicit values for the variables (C_j, $y_j(m)$ and x_j where $j = 0, 1$ and $m = 0, 1$) in function of the parameters (α, μ and λ). The cumulative distribution function (cdf) equals

$$\begin{aligned} S(t) &= \int_0^t s(u) du \\ &= \sum_{j=0}^{1} \sum_{m=0}^{1} c_j(m) \frac{1 - e^{\lambda\left(1-\frac{1}{x_j}\right)t}}{\lambda\left(1-\frac{1}{x_j}\right)}. \end{aligned} \tag{3.45}$$

And the tail probability is given by

$$\begin{aligned} \text{Prob}[S > t] &= 1 - S(t) \\ &= 1 - \sum_{j=0}^{1} \sum_{m=0}^{1} c_j(m) \frac{1 - e^{\lambda\left(1-\frac{1}{x_j}\right)t}}{\lambda\left(1-\frac{1}{x_j}\right)}. \end{aligned} \tag{3.46}$$

3.2.6 Discussion of results and numerical examples

In this subsection, we discuss the results obtained in the previous subsections, from a quantitative and a qualitative perspective, by means of some numerical examples.

Reference systems

In some of the results we compare our system with two "extreme" queueing systems which we call "worst" and "best". These are two boundaries for our system (lower and upper). Worst is a well-known queueing system ($M \mid M \mid 1$) with an infinite waiting room, *one* type of customers and *one* server with a mean service time of μ. It is clear that this is a lower boundary for the system, since there is always at least one server working in the system, when there are customers in the system. Best is another well-known queueing system ($M \mid M \mid 2$) with an infinite waiting room, *one* type of customers and *two* servers with a mean service time of μ. This is an upper boundary for the system, since at most two servers are working in the system. The solutions of both these queueing systems are standard and can be found in many books about queueing theory e.g. [6].

Impact of class clustering (parameter α)

The first interesting result obtained is the stability condition (3.5) which shows that the maximum achievable throughput of this system, expressed in average amount of work per time unit, is very directly determined by the degree of class clustering in the arrival process, as described by the cluster parameter α. From the stability condition, we can already deduce that the achievable throughput decreases with the cluster parameter α. It is also interesting to look at the extrema. For the first extremum, $\alpha = 1$, only one type of customers arrives and only one of the servers is being used. The system behaves as the system $M \mid M \mid 1$ and the throughput never exceeds 1 time unit of work per time unit. For the second extremum, $\alpha = 0$, the types of customers arrive alternating. Note that the throughput cannot exceed $\frac{3}{2}$ time units of work per time unit instead of the (possibly expected) 2 time units of work per time unit, which is the maximum throughput of the system $M \mid M \mid 2$. Thus even in the optimal case of $\alpha = 0$, both servers are not working all the time when the system would be constantly provided with new customers. In other words, the system is also in this case non-work-conserving. The reason for this is as follows: even when the customers arrive with alternating types, it is possible that the second customer (at the front of the system) has completed his service before the first customer has, since the second customer can have a shorter service time (the second customer overtakes the first customer). The third customer (now becoming the second) then still has to wait for service because the first customer occupies his server. The third customer then blocks the fourth customer that otherwise could have been served because his server is idle, ... So, the randomness of the service times is the culprit here.

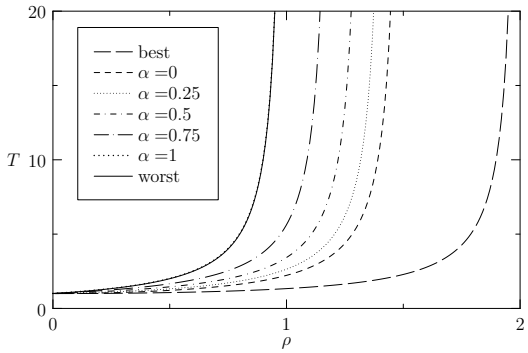


Figure 3.2: Mean system delay versus parameter ρ for a given service rate of 1

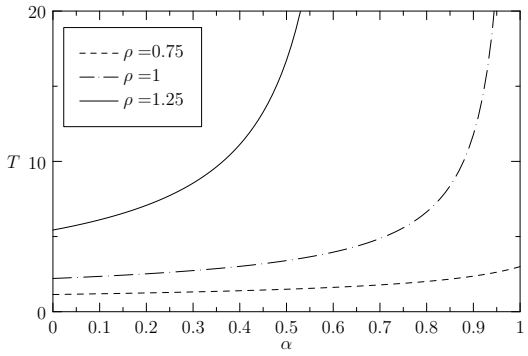


Figure 3.3: Mean system delay versus cluster parameter α for a given service rate of 1

We now move to the system delay and system occupancy in stable systems. Fig. 3.2 shows the mean system delay versus parameter ρ for different values of the cluster parameter α. The figure illustrates the (negative) impact of global FCFS, even in the best case (the types of customers arrive alternating or $\alpha = 0$), the system performs much "worse" than the case where the system would be work-conserving ($M \mid M \mid 2$). The figure also confirms some previously made observations. Our system behaves identical to the system $M \mid M \mid 1$ for $\alpha = 1$. When α increases, the stability region shrinks. When $\alpha = 0.5$, the type of the next customer is independent of the previous customer, and we can clearly see that neglecting the clustering in the arrival stream causes a considerable underestimation or overestimation of the performance of the system.

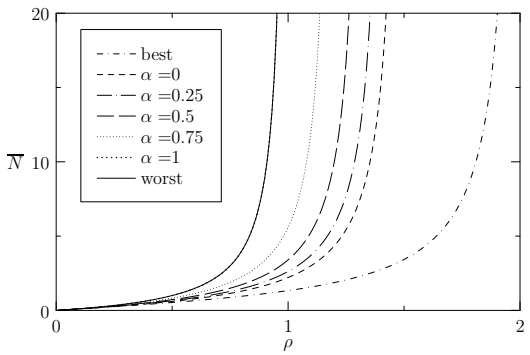


Figure 3.4: Mean system occupancy versus parameter ρ

Impact of load (parameter ρ)

In Fig. 3.3, the mean system delay is shown versus cluster parameter α for different values of the parameter ρ. Here, we notice that the impact of the cluster parameter is negligible for small values of ρ. This is not surprising. For small values of ρ, it is of lesser importance whether only one or both servers work since one server suffices to handle the incoming work. For bigger ρ, it is a different story. Here the cluster parameter has a big impact and can even lead to unstable systems. This illustrates that the cluster parameter should not be neglected.

In Fig. 3.4, the mean system occupancy is plotted versus parameter ρ for different values of the cluster parameter α. Here, similar conclusions are drawn as for the mean system delay (Fig. 3.2).

Use for dimensioning purposes

Fig. 3.5 represents the tail probability of the system delay for a given arrival and service rate of 1. The tail probability gives us the percentage of customers with a system delay greater than t units of time which can for example be of great importance for applications where the delay plays an important role. For example, given an arrival and service rate of 1 and cluster parameter of 0.75, circa 0.01% of the customers has a system delay greater than 51 units of time. This can be used for instance for dimensioning purposes. Note that the cluster parameter α should be measured, as this parameter has a crucial impact on performance. Notice also in Fig. 3.5 that the system is unstable for a given cluster parameter of 1.

Finally, Fig. 3.6 shows the tail probability of the system occupancy for a given arrival and service rate of 1. The tail probability can be considered as an approximate value for the loss probability in a system with finite storage

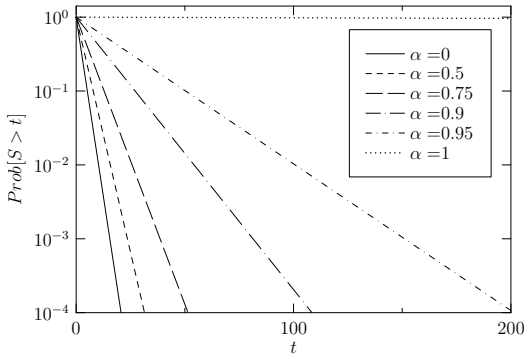


Figure 3.5: Tail probability of the system delay for a given arrival and service rate of 1

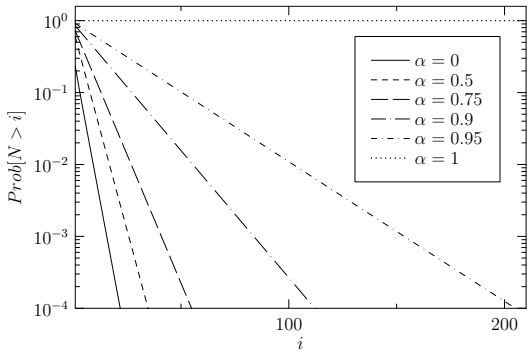


Figure 3.6: Tail probability of the system occupancy for a given arrival and service rate of 1

capacity equal to i places which can be used for dimensioning purposes. For example, given an arrival and service rate of 1 and cluster parameter of 0.75, the required buffer size is 56 for a loss ratio of 10^{-4}. Notice in Fig. 3.6 that the system is unstable for a given cluster parameter of 1 and the loss ratio of 10^{-4} is not achievable.

3.3 Two cluster parameters

In this section, we extend the model of the previous section by adding an extra cluster parameter. This way we can make a distinction between both types of customers and study the concept of class clustering to a greater extent.

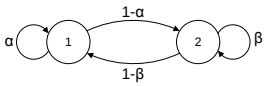


Figure 3.7: 2-state Markov chain to determine the type of an arriving customer

3.3.1 Mathematical model

We consider a continuous-time queueing model with infinite waiting room and two servers. Customers have exponential service times; server 1 serves customers with a service rate of μ_1 and server 2 serves customers with a service rate of μ_2. The servers are dedicated to a given class of customers. Server 1 only serves customers of one type (say type 1) and server 2 serves customers of the other type (type 2). The customers are served in their order of arrival, regardless of the class they belong to (gFCFS).

The customers enter the system according to a Poisson arrival process with arrival rate λ. The type of the arriving customer is determined by a two-state Markov chain (see Fig. 3.7). If the previous customer is of type 1, then the customer is of type 1 with probability α and of type 2 with probability $(1-\alpha)$. If the previous customer is of type 2, then the current customer is of type 1 with probability $(1-\beta)$ and of type 2 with probability β. Notice here already that we can transform α and β in two other parameters σ and K that have a more intuitive meaning. The transformations from (α,β) to (σ,K) are

$$\sigma = \frac{1-\beta}{2-\alpha-\beta}, \tag{3.47}$$

$$K = \frac{1}{2-\alpha-\beta}, \tag{3.48}$$

and from (σ,K) to (α,β)

$$\alpha = 1 - \frac{1-\sigma}{K}, \tag{3.49}$$

$$\beta = 1 - \frac{\sigma}{K}. \tag{3.50}$$

The intuitive meaning behind the parameter σ is that it represents the relative frequency of the type 1 customers, i.e., the fraction of customers that are of type 1 (2) is σ ($1-\sigma$ respectively). The parameter K on the other hand gives a clear indication about the correlation. The parameter is directly proportional to the mean number of customers of the same type

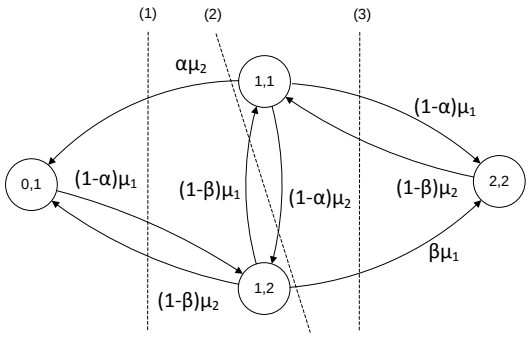


Figure 3.8: 4-state Markov chain to determine the stability condition

that arrive back-to-back. More specifically, we have

$$\begin{aligned} \mathrm{E}\left[\text{number of customers of type 1 arriving back-to-back}\right] &= \frac{1}{1-\alpha} \qquad (3.51)\\ &= \frac{K}{1-\sigma}, \\ \mathrm{E}\left[\text{number of customers of type 2 arriving back-to-back}\right] &= \frac{1}{1-\beta} \qquad (3.52)\\ &= \frac{K}{\sigma}, \end{aligned}$$

where $\mathrm{E}[\cdots]$ represents the expected value of the quantity between brackets. Notice here that when K equals 1, the types of consecutive customers in the arrival stream are uncorrelated and the model simplifies to that of Chapter 2.

3.3.2 Stability condition

When deriving the stability condition, we can presume that the system is constantly provided with new customers and the system will therefore be filled with at least 2 customers all the time. Note that we are only interested in the class of the customers in the set of leading customers, i.e., the 2 oldest customers in the system (possibly being served). These observations lead to the 4-state Markov chain depicted in Fig. 3.8. In state (m, t), m customers are of type 2 (and thus $2-m$ are of type 1) and the last customer in the set of leading customers has type t. Notice that we do not have states $(0, 2)$ and $(2, 1)$ since the last customer cannot be of type 2 (1) if all leading customers are of type 1 (2).

If we define $p(m, t)$ as the steady-state probabily to be in state (m, t), then we end up with the following balance equations (corresponding to the

dotted lines (1) to (3) in Fig. 3.8):

$$(1-\alpha)\mu_1 p(0,1) = \alpha\mu_2 p(1,1) + (1-\beta)\mu_2 p(1,2), \tag{3.53}$$

$$\mu_2 p(1,1) = \mu_1 p(1,2), \tag{3.54}$$

$$(1-\beta)\mu_2 p(2,2) = (1-\alpha)\mu_1 p(1,1) + \beta\mu_1 p(1,2). \tag{3.55}$$

These balance equations combined with the normalization condition

$$\sum_{m=0}^{2} (p(m,1) + p(m,2)) = 1, \tag{3.56}$$

where $p(0,2) = p(2,1) = 0$ by definition, yields

$$p(0,1) = \frac{\mu_2^2(1-\beta)(\alpha\mu_1 + (1-\beta)\mu_2)}{(1-\alpha)^2\mu_1^3 + (1-\alpha)\mu_1^2\mu_2 + (1-\beta)\mu_1\mu_2^2 + (1-\beta)\mu_2^3}, \tag{3.57}$$

$$p(1,1) = \frac{\mu_1^2\mu_2(1-\beta)(1-\alpha)}{(1-\alpha)^2\mu_1^3 + (1-\alpha)\mu_1^2\mu_2 + (1-\beta)\mu_1\mu_2^2 + (1-\beta)\mu_2^3}, \tag{3.58}$$

$$p(1,2) = \frac{\mu_1\mu_2^2(1-\beta)(1-\alpha)}{(1-\alpha)^2\mu_1^3 + (1-\alpha)\mu_1^2\mu_2 + (1-\beta)\mu_1\mu_2^2 + (1-\beta)\mu_2^3}, \tag{3.59}$$

$$p(2,2) = \frac{\mu_1^2(1-\alpha)((1-\alpha)\mu_1 + \beta\mu_2)}{(1-\alpha)^2\mu_1^3 + (1-\alpha)\mu_1^2\mu_2 + (1-\beta)\mu_1\mu_2^2 + (1-\beta)\mu_2^3}. \tag{3.60}$$

Having obtained the $p(m,t)$'s, we can now move on to the stability condition. Therefore, we postulate that the average amount of work per time unit that enters the system (ρ) is smaller than the average amount of work the system can execute per time unit, i.e., the average amount of work the system would execute per time unit when it would be constantly provided with new customers. Here, the system is able to execute 2 units of work per unit of time when both servers are able to work (when the system is in the state (1,1) or (1,2)). The system is able to execute 1 unit of work per unit of time when only one server is able to work (when the system is in state (0,1) or (2,2)). The stability condition is thus

$$\rho < p(0,1) + 2(p(1,1) + p(1,2)) + p(2,2), \tag{3.61}$$

or after using expressions (3.57) to (3.60)

$$\rho < \frac{(1 + \frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2})\left((1-\alpha)\frac{\mu_1}{\mu_2} + (1-\beta)\frac{\mu_2}{\mu_1} + 1\right)}{\frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2}\left((1-\alpha)\frac{\mu_1}{\mu_2} + 1\right) + (1-\beta)\frac{\mu_2}{\mu_1} + 1}, \tag{3.62}$$

where ρ (average amount of work that enters the system) is defined as follows

$$\rho = \rho_1 + \rho_2 \triangleq \frac{\sigma\lambda}{\mu_1} + \frac{(1-\sigma)\lambda}{\mu_2}. \tag{3.63}$$

To make the numerical results in section 3.3.5 more intuitive we use the transformations from equations (3.49) and (3.50). Here, we also already see that not the exact values of μ_1 and μ_2 are of importance but only the ratio. The stability condition becomes

$$\rho < \frac{(1 + \frac{1-\sigma}{\sigma}\frac{\mu_1}{\mu_2})\left(\frac{1-\sigma}{K}\frac{\mu_1}{\mu_2} + \frac{\sigma}{K}\frac{\mu_2}{\mu_1} + 1\right)}{\frac{1-\sigma}{\sigma}\frac{\mu_1}{\mu_2}\left(\frac{1-\sigma}{K}\frac{\mu_1}{\mu_2} + 1\right) + \frac{\sigma}{K}\frac{\mu_2}{\mu_1} + 1}. \tag{3.64}$$

3.3.3 System state diagram and balance equations

The system can be described by a continuous-time Markov chain where the state of the system is described by the triple (n, m, t). Here, n represents the number of customers in the system, m represents the number of customers of type 2 in the set of leading customers and t represents the type of the last customer in this set of customers (1 or 2). Notice that we do not have states $(n, 0, 2)$ and $(n, 2, 1)$ for $n > 1$ since the last customer cannot be of type 2 (1) if all leading customers are of type 1 (2). Remark the need for states $(1, 0, 2)$ and $(1, 1, 1)$. It is possible that the last customer has already left the system and thus has overtaken the customer still in the system (by having a shorter service time). The remaining customer is not necessarily the customer that arrived last. State $(0, t)$ represents the empty system where the last customer that arrived is of type t. This is thus a QBD process (see also [7]) with four phases and the levels are represented by the number of customers in the system.

If we define $p(n, m, t)$ as the steady-state probability to be in state (n, m, t) (and $p(0, t)$ to be in state $(0, t)$), we end up with the following balance and boundary equations (observe transitions to and from states (1)-(14) in Fig. 3.9):

$$\lambda p(0,1) = \mu_1 p(1,0,1) + \mu_2 p(1,1,1), \tag{3.65}$$

$$\lambda p(0,2) = \mu_1 p(1,0,2) + \mu_2 p(1,1,2), \tag{3.66}$$

$$\begin{aligned}(\lambda + \mu_1) p(1,0,1) = {} & \mu_1 p(2,0,1) + \mu_2 p(2,1,1) \\ & + \lambda(\alpha p(0,1) + (1-\beta) p(0,2)),\end{aligned} \tag{3.67}$$

$$(\lambda + \mu_1) p(1,0,2) = \mu_2 p(2,1,2), \tag{3.68}$$

$$(\lambda + \mu_2) p(1,1,1) = \mu_1 p(2,1,1), \tag{3.69}$$

$$\begin{aligned}(\lambda + \mu_2) p(1,1,2) = {} & \mu_1 p(2,1,2) + \mu_2 p(2,2,2) \\ & + \lambda((1-\alpha) p(0,1) + \beta p(0,2)),\end{aligned} \tag{3.70}$$

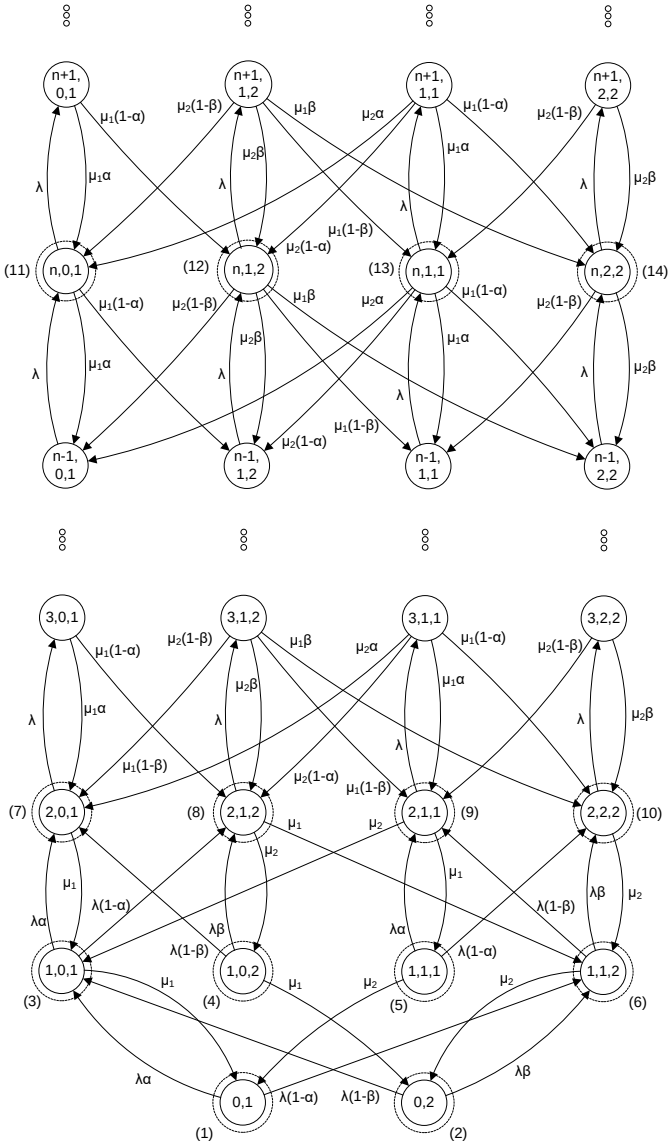


Figure 3.9: State Diagram

$$
\begin{aligned}
(\lambda+\mu_1)p(2,0,1) &= \mu_1\alpha p(3,0,1) + \mu_2((1-\beta)p(3,1,2) + \alpha p(3,1,1)) \\
&\quad + \lambda(\alpha p(1,0,1) + (1-\beta)p(1,0,2)), && (3.71)\\
(\lambda+\mu_1+\mu_2)p(2,1,2) &= \mu_1(1-\alpha)p(3,0,1) \\
&\quad + \mu_2(\beta p(3,1,2) + (1-\alpha)p(3,1,1)) \\
&\quad + \lambda((1-\alpha)p(1,0,1) + \beta p(1,0,2)), && (3.72)\\
(\lambda+\mu_1+\mu_2)p(2,1,1) &= \mu_1((1-\beta)p(3,1,2) + \alpha p(3,1,1)) \\
&\quad + \mu_2(1-\beta)p(3,2,2) \\
&\quad + \lambda(\alpha p(1,1,1) + (1-\beta)p(1,1,2)), && (3.73)\\
(\lambda+\mu_2)p(2,2,2) &= \mu_1(\beta p(3,1,2) + (1-\alpha)p(3,1,1)) + \mu_2\beta p(3,2,2) \\
&\quad + \lambda((1-\alpha)p(1,1,1) + \beta p(1,1,2)), && (3.74)\\
(\lambda+\mu_1)p(n,0,1) &= \mu_1\alpha p(n+1,0,1) \\
&\quad + \mu_2((1-\beta)p(n+1,1,2) + \alpha p(n+1,1,1)) \\
&\quad + \lambda p(n-1,0,1), \quad n \geq 3, && (3.75)\\
(\lambda+\mu_1+\mu_2)p(n,1,2) &= \mu_1(1-\alpha)p(n+1,0,1) \\
&\quad + \mu_2(\beta p(n+1,1,2) + (1-\alpha)p(n+1,1,1)) \\
&\quad + \lambda p(n-1,1,2), \quad n \geq 3, && (3.76)\\
(\lambda+\mu_1+\mu_2)p(n,1,1) &= \mu_1((1-\beta)p(n+1,1,2) + \alpha p(n+1,1,1)) \\
&\quad + \mu_2(1-\beta)p(n+1,2,2) \\
&\quad + \lambda p(n-1,1,1), \quad n \geq 3, && (3.77)\\
(\lambda+\mu_2)p(n,2,2) &= \mu_1(\beta p(n+1,1,2) + (1-\alpha)p(n+1,1,1)) \\
&\quad + \mu_2\beta p(n+1,2,2) \\
&\quad + \lambda p(n-1,2,2)), \quad n \geq 3. && (3.78)
\end{aligned}
$$

For example, the left-hand side of equation (3.76) represents the system leaving state $(n,1,2)$ with rate λ (a new customer enters the system), rate μ_1 (a customer of type 1 leaves the system) and rate μ_2 (a customer of type 2 leaves the system). The right-hand side of the equation is a bit more involved. We go to state $(n,1,2)$ in four cases. First, with rate λ (an arrival occurs) state $(n,1,2)$ is reached from state $(n-1,1,2)$. The arriving customer does not change the leading customers since at least 2 customers are already present when the customer arrives ($n \geq 3$). Secondly, the system goes from state $(n+1,0,1)$ to state $(n,1,2)$ with rate $\mu_1(1-\alpha)$. This happens when a customer of type 1 leaves the system and the "new" customer in the set of leading customers is of type 2 (with probability $1-\alpha$ since the previous last customer of the leading customers was of type 1). Analogously, the system can go with rate $\mu_2\beta$ from state $(n+1,1,2)$ to state $(n,1,2)$ and with rate $\mu_2(1-\alpha)$ from state $(n+1,1,1)$ to state $(n,1,2)$.

3.3.4 Analysis of the distribution and moments of the system occupancy

To tackle the analysis of the system occupancy, we make use of probability generating functions (pgf). First, the pgf of the system occupancy is determined, already giving us straightforwardly some important performance measures (e.g. mean system occupancy). We will also invert the obtained pgfs using partial fraction expansion to obtain (tail asymptotics of) the probability mass function (pmf).

Relation between pgf P(z) and some partial pgfs

The pgf of the (total) number of customers in the system is given by

$$\begin{aligned} P(z) = p(0) + z \cdot (p(1,0,1) + p(1,0,2) + p(1,1,1) + p(1,1,2)) \\ + Q_0(z) + Q_1(z) + Q_2(z) + Q_3(z), \end{aligned} \tag{3.79}$$

where we introduce the partial pgfs

$$Q_0(z) \triangleq \sum_{n=2}^{\infty} p(n,0,1) z^n, \tag{3.80}$$

$$Q_1(z) \triangleq \sum_{n=2}^{\infty} p(n,1,2) z^n, \tag{3.81}$$

$$Q_2(z) \triangleq \sum_{n=2}^{\infty} p(n,1,1) z^n, \tag{3.82}$$

$$Q_3(z) \triangleq \sum_{n=2}^{\infty} p(n,2,2) z^n. \tag{3.83}$$

Determination of the partial pgfs

Equations (3.75) to (3.78) are multiplied by z^n and summed over all $n \geqslant 3$. We find

$$\begin{aligned} &(\lambda + \mu_1)(Q_0(z) - z^2 p(2,0,1)) = \\ &\quad \frac{1}{z} \left[\mu_1 \alpha (Q_0(z) - z^3 p(3,0,1) - z^2 p(2,0,1)) \right. \\ &\quad + \mu_2((1-\beta)(Q_1(z) - z^3 p(3,1,2) - z^2 p(2,1,2)) \\ &\quad \left. + \alpha(Q_2(z) - z^3 p(3,1,1) - z^2 p(2,1,1))) \right] \\ &\quad + \lambda z Q_0(z), \end{aligned} \tag{3.84}$$

$$
\begin{aligned}
&(\lambda+\mu_1+\mu_2)(Q_1(z)-z^2p(2,1,2)) = \\
&\quad \frac{1}{z}\big[\mu_1(1-\alpha)(Q_0(z)-z^3p(3,0,1)-z^2p(2,0,1)) \\
&\quad +\mu_2(\beta(P_1(z)-z^3p(3,1,2)-z^2p(2,1,2)) \\
&\quad +(1-\alpha)(Q_2(z)-z^3p(3,1,1)-z^2p(2,1,1)))\big] \\
&\quad +\lambda z Q_1(z), \qquad (3.85)\\
&(\lambda+\mu_1+\mu_2)(Q_2(z)-z^2p(2,1,1)) = \\
&\quad \frac{1}{z}\big[\mu_1((1-\beta)(Q_1(z)-z^3p(3,1,2)-z^2p(2,1,2)) \\
&\quad +\alpha(Q_2(z)-z^3p(3,1,1)-z^2p(2,1,1))) \\
&\quad +\mu_2(1-\beta)(Q_3(z)-z^3p(3,2,2)-z^2p(2,2,2))\big] \\
&\quad +\lambda z Q_2(z), \qquad (3.86)\\
&(\lambda+\mu_2)(Q_3(z)-z^2p(2,2,2)) = \\
&\quad \frac{1}{z}\big[\mu_1(\beta(Q_1(z)-z^3p(3,1,2)-z^2p(2,1,2)) \\
&\quad +(1-\alpha)(Q_2(z)-z^3p(3,1,1)-z^2p(2,1,1))) \\
&\quad +\mu_2\beta(Q_3(z)-z^3p(3,2,2)-z^2p(2,2,2))\big] \\
&\quad +\lambda z Q_3(z)). \qquad (3.87)
\end{aligned}
$$

Determination of the pgf

We now calculate $P(z)$ from (3.79). If we solve the set of linear equations (3.65) to (3.74) and (3.84) to (3.87) and insert the solutions in (3.79) this equation translates into an equation that only contains known quantities, except for four unknown probabilities in the numerator. These can be determined, in general, by invoking the well-known property that pgfs such as $P(z)$ are bounded inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z-plane, at least when the stability condition (3.61) of the queueing system is met (only in such a case our analysis was justified and $P(z)$ can be viewed as a legitimate pgf). Now, it can be shown by means of Rouché's theorem from complex analysis [8, 9] that the denominator of (3.79) has exactly four zeroes inside the closed unit disk of the complex z-plane, one of which is equal to 1, as soon as the stability condition (3.61) is fulfilled. It is clear that these four zeroes should also be zeroes of the numerator of (3.79), as $P(z)$ must remain bounded in those points. We conclude with the calculation of the three remaining zeroes that are inside the closed unit disk, using numerical methods. For the zeroes inside the closed unit disk (z_0, z_1, z_2), the requirement that the numerator should vanish yields three linear equations for the four unknowns. For the zero $z = 1$, this condition is fulfilled regardless of the values of the unknowns, since the numerator

of (3.79) contains a factor $z-1$. A fourth linear equation can however be obtained by invoking the normalizing condition of the pgf $P(z)$, i.e., the condition $P(1)=1$. In general, the four unknown probabilities can be found as the solutions of the four established linear equations. Substitution of the obtained values in (3.79) then leads to a fully determined expression for the steady-state pgf $P(z)$ of the system occupancy. $P(z)$ is a rational function (the quotient of two polynomials of degree 4).

From this result, various performance measures of practical importance can then be derived. For instance, the mean system occupancy can be found as $\overline{N}=P'(1)$. The mean system delay T can then be calculated using Little's Law [10].

$$T=\frac{\overline{N}}{\lambda}. \tag{3.88}$$

3.3.5 Discussion of results and numerical examples

In this section, we discuss the results obtained in the previous sections, both from a qualitative perspective and by means of some numerical examples. Before discussing the results, we introduce two new parameters

$$\omega \triangleq \frac{\frac{\sigma}{\mu_1}}{\frac{\sigma}{\mu_1}+\frac{1-\sigma}{\mu_2}}=\frac{\rho_1}{\rho_1+\rho_2}, \tag{3.89}$$

$$d \triangleq \frac{\mu_1}{\mu_1+\mu_2}. \tag{3.90}$$

These parameters will allow us to interpret the results more intuitively. The parameter ω represents the relative load of customers of type 1 and d represents the relative service rate of type 1.

Impact of class clustering (parameter K)

Fig. 3.10 shows ρ_{sup}, least upper bound of the set of values ρ where the system is stable versus parameter ω, with $\mu_1=20$ and $\mu_2=1$. It is clear that the system where customers have the tendency to arrive back-to-back (higher K) performs "worse" than the system where customers have the tendency to arrive more alternatingly (smaller K). An observation that is also confirmed in Fig. 3.11 where the mean system occupancy versus parameter ρ with $\sigma=\frac{1}{2}$, $\mu_1=1$ and $\mu_2=20$ (and thus $\omega=\frac{20}{21}$ and $d=\frac{1}{21}$) is shown. Those figures illustrate that it is not possible to ignore the concept of class clustering for our system.

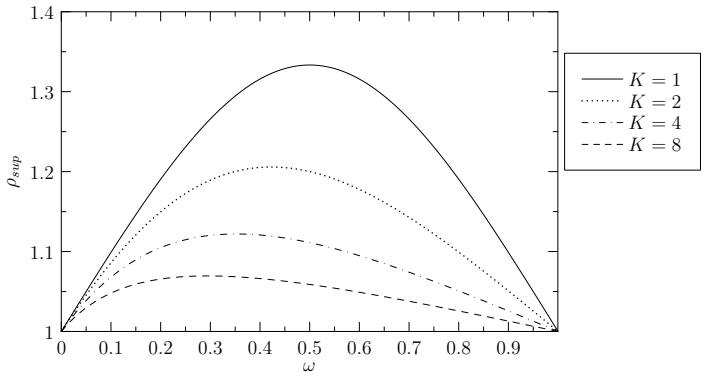


Figure 3.10: ρ_{sup}, the least upper bound of the set of values ρ where the system is stable versus parameter ω, with $\mu_1 = 20$ and $\mu_2 = 1$

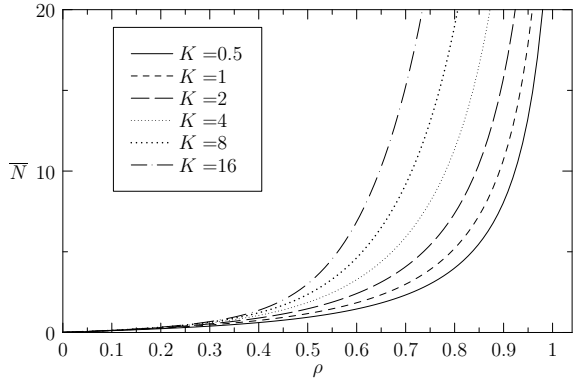


Figure 3.11: Mean system occupancy versus parameter ρ, with $\sigma = \frac{1}{2}$, $\mu_1 = 1$ and $\mu_2 = 20$ ($\omega = \frac{20}{21}$ and $d = \frac{1}{21}$)

Impact of the load and service rate balance between customers of type 1 and customers of type 2 (parameters ω and d)

In Fig. 3.10, we notice that the maximum achievable throughput when $K = 1$ is obtained for a perfectly balanced system ($\omega = \frac{1}{2}$). However, the more class clustering (K increases), the more this maximum will move towards a situation where the fastest server gets a higher relative load. This might be a little contra-intuitive. In a system without blocking, the maximum achievable throughput is achieved when our system is perfectly balanced (both servers get a load of 1, or a total load of 2) irrespective of K. In the system with blocking, this maximum achievable throughput lies between 1 (the load one server can process) and 2 (the load two servers can process), as can also be seen from Fig. 3.10. The exact maximum is actually determined

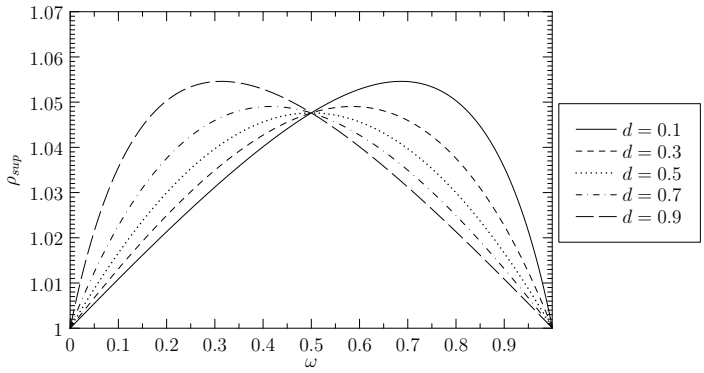


Figure 3.12: ρ_{sup}, the least upper bound of the set of values ρ where the system is stable versus parameter ω, with $K = 10$ and $\mu_2 = 1$

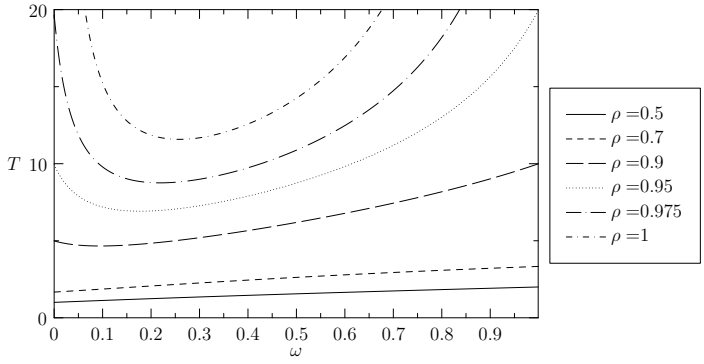


Figure 3.13: Mean system time versus parameter ω, with $K = 5$, $\mu_1 = 1$ and $\mu_2 = 2$ ($d = \frac{1}{3}$)

by maximizing the fraction of time both servers work simultaneously, i.e., when the two leading customers are of opposite type. This observation is also confirmed by Fig. 3.12, which represents ρ_{sup}, the least upper bound of the set of values ρ where the system is stable versus parameter ω, with $K = 10$ and $\mu_2 = 1$. We notice here that it is not always ideal to have a symmetric system (where the workload is equally balanced and both servers have the same service rate). Even more surprising is that when there is negative correlation in the types of consecutively arriving customers ($K < 1$), the slowest server should get a higher relative load if physically possible.

In Fig. 3.13, the mean system time versus parameter ω, with $K = 5$, $\mu_1 = 1$ and $\mu_2 = 2$ ($d = \frac{1}{3}$) is shown. We see that the mean system time is rather independent of the load balance (ω), for small total load (ρ). The load balance (whether or not the slowest server is also working) becomes

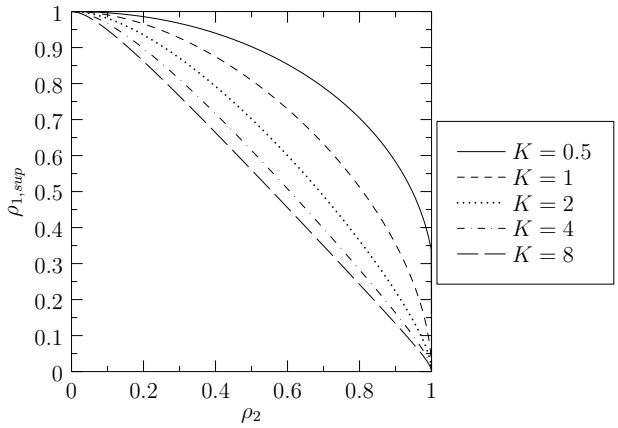


Figure 3.14: $\rho_{1,sup}$, the least upper bound of the set of values ρ_1 where the system is stable versus parameter ρ_2, with $\sigma = \frac{1}{2}$

only of importance when the total load becomes too much for one server to handle (when ρ approaches 1). From Fig. 3.13, it can be observed that it is even better for our system to have only one type of customer (of the fastest server) for a small load. This, however, is mainly due to the fact that the service time is included in the system time.

Impact of customers of one type on customers of the other type

Fig. 3.14 represents $\rho_{1,sup}$, the least upper bound of the set of values ρ_1 where the system is stable versus parameter ρ_2, with $\sigma = \frac{1}{2}$. The more class clustering (higher K), the more the customers of different types have an influence on each other. This is intuitively also clear since the more class clustering (or more customers of the same type arriving consecutively), the more a customer of a different type will be blocked behind the group of customers of the same type.

References

[1] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* 1995.

[2] Ivo Adan, Ton de Kok, and Jacques Resing. *A Multi-Server Queueing Model With Locking.* EJOR, 116:16–26, 2000.

[3] Ivo Adan, Jaap Wessels, and Henk Zijm. *A Compensation Approach for Two-Dimensional Markov Processes.* Advances in Applied Probability, 25:783–817, 1993.

[4] Dimitris Bertsimas. *An Exact FCFS Waiting Time Analysis for a General Class of G/G/s Queueing Systems.* Queueing Systems, pages 305–320, 1988.

[5] Dimitris Bertsimas. *An Analytic Approach to a General Class of G/G/s Queueing Systems.* Operations Research, pages 139–155, 1990.

[6] Pavel Petrovich Bocharov and Ciro D'Apice. *Queueing Theory.* 2004.

[7] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* 1981.

[8] Mario O. Gonzáles. *Classical Complex Analysis.* 1992.

[9] Herwig Bruneel and Byung Guk Kim. *Discrete-time Models for Communication Systems Including ATM.* 1993.

[10] Leonard Kleinrock. *Theory, Volume 1, Queueing Systems.* 1975.

4

The impact of the global First-Come-First-Served scheduling with presorting

4.1 Introduction

In this chapter, we tackle the second objective of this dissertation. This objective is to have a better grasp on the concept of gFCFS service discipline with presorting, i.e., all arriving customers are accommodated in one single FCFS queue, regardless of their type, with an exception of the first P customers. For the first P customers the FCFS rule holds only within the type, i.e., customers of different types can overtake each other in order to be served. This models the concept of a turn lane. The result of the work in previous chapters can in fact be regarded as a worst case scenario (no turn lane), while two separate queues (infinite turn lane) can be seen as a best case scenario.

The structure of the chapter is as follows: we first describe the mathematical model in Section 4.2. In Section 4.3, we briefly return to the problem of the stability condition. Next, in Section 4.4, we analyse the distribution of the number of customers in the system in two steps. The chapter continues with a discussion of the results and some numerical examples in Section 4.5.

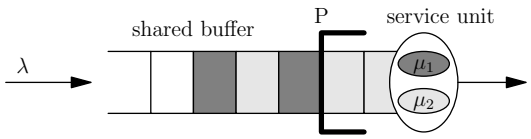


Figure 4.1: Model of the system with global FCFS and presorting

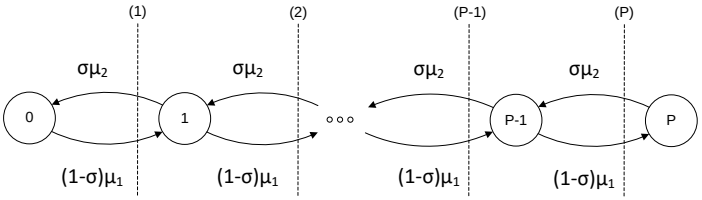


Figure 4.2: $(P+1)$-state Markov chain to determine the stability condition

4.2 Mathematical model

We consider a continuous-time queueing system (as shown in Fig. 4.1) with infinite storage capacity. The customers enter the system according to a Poisson arrival process with mean arrival rate λ. The types of consecutive customers are independent, i.e., an arriving customer is of type 1 with probability σ and of type 2 with probability $1-\sigma$. Two types of customers (type 1 and 2) are to be served by two dedicated servers (server 1 and 2). Customers of type 1 (2) are served by server 1 (2) and have an exponential service time with a service rate of μ_1 (μ_2). All service times are independent. The system operates under the gFCFS policy with presorting (P-gFCFS), such that the customers are served in the order of their arrival, regardless of the class they belong to, except for the P leading customers, i.e., the P oldest customers in the system (those being served included). The leading customers are being served according to a *per-type* FCFS policy. Obviously, the most important consequence is that server 1 (2) is working if there is at least one customer of type 1 (2) in the leading customers. This also means that server 1 (2) is not working if all leading customers are of type 2 (1) even though there is a customer of type 1 (2) in the system. This queueing system can be considered as an appropriate model for the presorting lane. We briefly explain why in Section 4.5.

4.3 Stability condition

The system is stable when the average amount of work per time unit that enters the system (ρ) is smaller than the average amount of work the system

can execute per time unit, i.e., the average amount of work the system would execute per time unit when it would be constantly provided with new customers. Here we can define ρ as the average amount of work of type 1 and 2 per unit time

$$\rho = \rho_1 + \rho_2 \triangleq \frac{\sigma\lambda}{\mu_1} + \frac{(1-\sigma)\lambda}{\mu_2}. \tag{4.1}$$

To determine the average amount of work the system would execute per time unit, we first calculate the steady-state probabilities to be in states where either one or both servers are working. Some observations can help us to construct a Markov chain to calculate these probabilities. First of all, since we are looking at the stability condition, we can presume that the system is constantly provided with new customers and the system will therefore be filled with at least P customers all the time. Second, only the leading customers are of importance since customers can only be served when they are in the first P customers of the system because of the P-gFCFS service discipline. The exact queueing order of the leading customers is also of no importance; once a customer is one of the leading customers, he can be served by his server if no other customers of his type are in front of him. Therefore, we are only interested in the number of customers of type 1 and 2 in the leading customers. Notice here that when we know the number of customers of type 2 in the leading customers, we also know the number of customers of type 1 (so we only have to keep track of the number of customers of type 2 in the leading customers). These observations lead to the $(P+1)$-state Markov chain in Fig. 4.2. The state m represents that m leading customers are of type 2 (and therefore, $P-m$ of type 1). The rate to go from state m to state $m-1$ is $\sigma\mu_2$; namely a rate μ_2 to end the service in state m of the customer with type 2 multiplied with the probability σ that the new P-th customer of our system is of type 1. Similarly, the rate to go from state m to state $m+1$ is $(1-\sigma)\mu_1$. It is clear that Fig. 4.2 models the well-known birth-and-death process for a $M|M|1|P$ queue [1] with arrival rate $(1-\sigma)\mu_1$ and service rate $\sigma\mu_2$. The probability $p(m)$ to be in state m is known to be given by

$$p(m) = \frac{\left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^m \left(1 - \frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)}{1 - \left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^{P+1}}. \tag{4.2}$$

The system is able to execute 1 unit of work per unit of time when only one server is able to work, i.e., when the system is in state 0 or state P. Otherwise, when both servers are working, the system executes 2 units of

work per unit of time. Therefore, the stability condition is

$$\rho < p(0) + 2\sum_{m=1}^{P-1} p(m) + p(P) \tag{4.3}$$

$$\rho < \frac{\left(1 + \frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)\left(1 - \left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^P\right)}{1 - \left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^{P+1}}. \tag{4.4}$$

Equation (4.4) can be rewritten as

$$\lambda < \frac{\left(\frac{\sigma}{\mu_1}\right)^P - \left(\frac{1-\sigma}{\mu_2}\right)^P}{\left(\frac{\sigma}{\mu_1}\right)^{P+1} - \left(\frac{1-\sigma}{\mu_2}\right)^{P+1}}. \tag{4.5}$$

which represents that on average, there are not more arrivals than service completions.

4.4 Analysis of the distribution and moments of the system occupancy

We now concentrate on the system occupancy distribution. With the observations of Section 4.3 in mind, the whole system can be described by a continuous-time Markov chain where the state of the system is characterised by a pair (n, m) where $n \geq 0, 0 \leq m \leq min(n, P)$. Here n represents the total number of customers in the system (those in service included) and m represents the number of leading customers that are of type 2. The Markov chain is thus a QBD process with $P + 1$ phases, and the levels are represented by the number of customers in the system. In QBD processes, the balance equations can be divided into boundary equations and repeating equations [2]. We will regard both separately.

4.4.1 Repeating equations

We start by looking at the repeating part of the Markov chain. QBD processes are commonly solved by using matrix-geometric techniques. Grassmann states in [3] that the problem with matrix-geometric methods is that they do not preserve the sparsity of the matrices involved. In other words, the matrix-geometric method does not exploit the fact that the matrices involved are tridiagonal which means that the computational effort can be reduced significantly. Although eigenvalues also have their problems, these

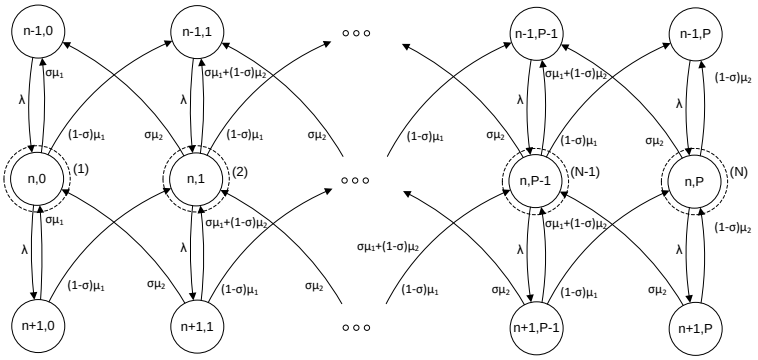


Figure 4.3: Repeating part of the QBD

seem to be manageable for the problem under investigation. In this chapter we will therefore use the method of eigenvalues as described in [3].

The repeating part of the QBD is shown in Fig. 4.3. The repeating equation can be written as

$$0 = \underline{\pi}_{n-1}Q_1 + \underline{\pi}_n Q_0 + \underline{\pi}_{n+1}Q_{-1}, n \geq P \tag{4.6}$$

with $\underline{\pi}_n = [\pi_{n,0}, \cdots, \pi_{n,P}]$ and $\pi_{n,m}$ represents the steady-state probability to be in state (n, m), for $m = 0, \cdots, P$ and $n \geq P$. From Fig. 4.3 these matrices are deduced as

$$Q_1 = \begin{bmatrix} \lambda & & \\ & \ddots & \\ & & \lambda \end{bmatrix}, \tag{4.7}$$

$$Q_0 = \begin{bmatrix} -\lambda-\mu_1 & & & & \\ & -\lambda-\mu_1-\mu_2 & & & \\ & & \ddots & & \\ & & & -\lambda-\mu_1-\mu_2 & \\ & & & & -\lambda-\mu_2 \end{bmatrix} \tag{4.8}$$

$$Q_{-1} = \begin{bmatrix} \sigma\mu_1 & (1-\sigma)\mu_1 & & & \\ \sigma\mu_2 & \sigma\mu_1+(1-\sigma)\mu_2 & (1-\sigma)\mu_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \sigma\mu_2 & \sigma\mu_1+(1-\sigma)\mu_2 & (1-\sigma)\mu_1 \\ & & & \sigma\mu_2 & (1-\sigma)\mu_2 \end{bmatrix} \tag{4.9}$$

QBD processes have a well known geometric relation, which means equation (4.6) has solutions of the form $\underline{d}x^n$ with $\underline{d}$ a vector of size $P+1$ (see also [2]). Replacing $\underline{\pi}_n$ by $\underline{d}x^n$ in (4.6), yields

$$0 = \underline{d}x^{n-1}Q_1 + \underline{d}x^n Q_0 + \underline{d}x^{n+1}Q_{-1}, \tag{4.10}$$

or by dividing by x^{n-1}

$$0 = \underline{d}Q(x), \tag{4.11}$$

where

$$Q(x) = Q_1 + Q_0 x + Q_{-1} x^2. \tag{4.12}$$

The row vector $\underline{d}$ is referred to as the eigenvector and the scalar x is called the eigenvalue, because they can be found by solving a so-called generalized eigenvalue problem (sometimes referred to as matrix pencil) given by (4.11) [4]. It was shown in [5] that if the process is recurrent, and all eigenvalues are distinct, there are $P+1$ distinct solutions of the form $\underline{d}x^n$. We denote the k-th couple (eigenvector, eigenvalue) by $(\underline{d}^{(k)}, x_k)$, $k = 0, \cdots, P$.

The problem at hand reduces to finding these eigenvalues and eigenvectors. Expansion of (4.11) yields

$$0 = d_0 \left[\lambda - (\lambda + \mu_1)x + \sigma\mu_1 x^2\right] + d_1 \left[\sigma\mu_2 x^2\right], \tag{4.13}$$

$$\begin{aligned} 0 = {} & d_{i-1} \left[(1-\sigma)\mu_1 x^2\right] + d_i \left[\lambda - (\lambda + \mu_1 + \mu_2)x + (\sigma\mu_1 + (1-\sigma)\mu_2)x^2\right] \\ & + d_{i+1} \left[\sigma\mu_2 x^2\right], i = 1, \cdots, P-1 \end{aligned} \tag{4.14}$$

$$0 = d_{P-1} \left[(1-\sigma)\mu_1 x^2\right] + d_P \left[\lambda - (\lambda + \mu_2)x + (1-\sigma)\mu_2 x^2\right]. \tag{4.15}$$

We now introduce functions $d_i(x)$ satisfying $d_i(x) = d_i$ whenever x is an eigenvalue or $det(Q(x)) = 0$. We can set $d_0(x) = d_0 = 1$ and replace $d_i = d_i(x)$ in (4.13) to (4.15) and solve for $d_i(x)$. This yields

$$d_1(x) = -\frac{\lambda - (\lambda + \mu_1)x + \sigma\mu_1 x^2}{\sigma\mu_2 x^2}, \tag{4.16}$$

$$\begin{aligned} d_{i+1}(x) = {} & -\frac{1}{\sigma\mu_2 x^2} \left(d_{i-1}(x) \left[(1-\sigma)\mu_1 x^2\right]\right. \\ & \left. + d_i(x) \left[\lambda - (\lambda + \mu_1 + \mu_2)x + (\sigma\mu_1 + (1-\sigma)\mu_2)x^2\right]\right), \\ & i = 1, \cdots, P-1, \end{aligned} \tag{4.17}$$

$$\begin{aligned} d_{P+1}(x) = {} & -\frac{1}{\sigma\mu_2 x^2} \left(d_{P-1}(x) \left[(1-\sigma)\mu_1 x^2\right]\right. \\ & \left. + d_P(x) \left[\lambda - (\lambda + \mu_2)x + (1-\sigma)\mu_2 x^2\right]\right). \end{aligned} \tag{4.18}$$

Notice here we have introduced $d_{P+1}(x)$ and as shown by Wilkinson in [4], $det(Q(x)) = d_{P+1}(x) \prod_{i=0}^{P} (-\sigma\mu_2 x^2)$. The problem then transforms in finding an x such that $d_{P+1} = d_{P+1}(x) = 0$ (and thus $det(Q(x)) = 0$). Essentially, to find the eigenvalues, we use the fact that $\{d_i(x), i = 0, 1, \cdots, P+1\}$ is a Sturm sequence. A Sturm sequence is any sequence with (i) $d_0(x)$ has no real roots (does not change its sign), (ii) $d_i(\epsilon) = 0$ implies $d_{i-1}(\epsilon)d_{i+1}(\epsilon) < 0$ ($sign(d_{i-1}(\epsilon)) = -sign(d_{i+1}(\epsilon))$), (iii) all real roots of $d_{P+1}(x)$ are simple [6]. Fundamental to Sturm sequences are sign

variations. The number of sign variations in the Sturm sequence $\{d_i(x), i = 0, 1, \cdots, P+1\}$ $(n(x))$ is given by

$$n(x) = \#\{d_i(x)d_{i+1}(x) < 0, 0 \leq i < P\} + \#\{d_i(x) = 0, 0 \leq i < P\}. \quad (4.19)$$

In [6], it is proved that there are at least $\mid n(x_1) - n(x_2) \mid$ eigenvalues between x_1 and x_2. In this specific case, $n(0+) = P+1$ and $n(1-) = 0$ or there are at least $P+1$ eigenvalues between $0+$ and $1-$. This means all $P+1$ eigenvalues within the unit circle are accounted for and we can use the divide-and-conquer algorithm described in [3], which is an extension of the binary search algorithm. This means we will recursively divide the search interval into two parts and discard any interval not containing an eigenvalue. We do this until we have found $P+1$ intervals containing at least one eigenvalue. Since there are only $P+1$ eigenvalues within the unit circle, all $P+1$ intervals will hold exactly one eigenvalue. Once we determined the intervals with only one eigenvalue, we can use the false position method to determine the exact value. After determining all the eigenvalues, equations (4.16), (4.17) and (4.18) can be used to determine the corresponding eigenvectors recursively.

Any linear combination of these solutions also forms a solution:

$$\underline{\pi}_n = \sum_{k=0}^{P} c_k \underline{d}^{(k)} x_k^n = \underline{c}\Lambda^n D, n \geqslant P \quad (4.20)$$

where $\underline{c} = [c_0, \cdots, c_P]$, $\Lambda = diag(x_k)$ and $D = \left[\underline{d}^{(0)}, \cdots, \underline{d}^{(P)}\right]^T$. Notice here that $\underline{c}$ can be determined by solving the boundary equations, which we will do next.

4.4.2 Boundary equations

The boundary states describe a QBD with maximum $P+1$ phases. The number of states in the boundary conditions is thus dependent on the parameter P. For example, when $P = 10$, we have 55 number of states in the boundary conditions, but when $P = 100$, we have already 5050 number of states. An efficient method for solving the boundary conditions is needed. The boundary states themselves form a level-dependent QBD. Only a few approaches found in literature try to exploit the specific structure in the level-dependent case. We will follow the algorithm described in [7] and [8]. The algorithm is based on matrix continued fractions (MCF).

First we determine the generator matrix Q for $\tilde{\underline{\pi}} = \left[\underline{\pi}_0, \cdots, \underline{\pi}_{P-1}, \underline{c}\right]$ of the Markov chain of the boundary conditions where $\underline{\pi}_n = [\pi_{n,0}, \cdots, \pi_{n,m}]$ and $\pi_{n,m}$ represents the steady-state probability to be in state (n, m), for

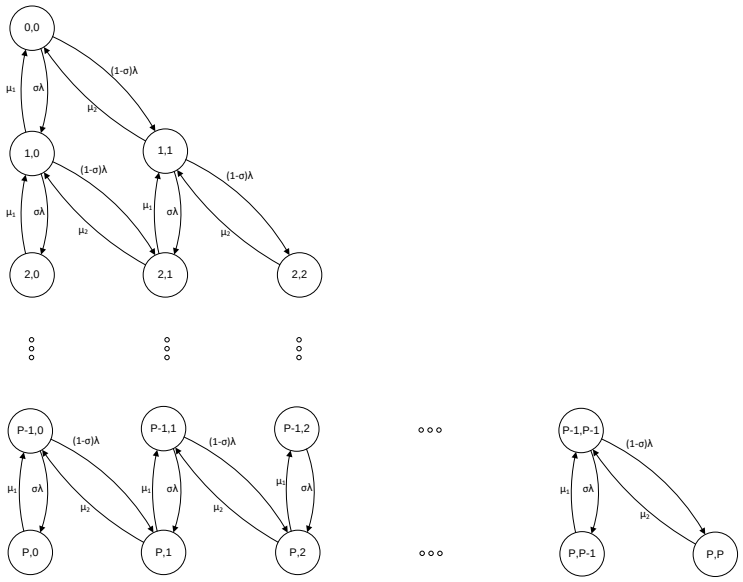


Figure 4.4: Boundary part of the QBD

$m = 0, \cdots, n$ and $n < P$. Notice here that we have replaced $\underline{\pi}_P$ by $\underline{c}$ because we are interested in the exact values of $\underline{c}$ (which are the only unknowns left in (4.20) at the end of Section 4.4.1). This generator matrix is given by

$$Q = \begin{bmatrix} Q_{0,0} & Q_{0,1} & & & & \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & & & \\ & Q_{2,1} & Q_{2,2} & Q_{2,3} & & \\ & \ddots & \ddots & \ddots & & \\ & & Q_{P-1,P-2} & Q_{P-1,P-1} & Q_{P-1,P} & \\ & & & \Lambda^P D Q_{P,P-1} & \Lambda^P D Q_{P,P} + \Lambda^{P+1} D Q_{-1} \end{bmatrix}. \tag{4.21}$$

where

$$Q_{i,i+1} = \begin{bmatrix} \sigma\lambda & (1-\sigma)\lambda & & \\ & \ddots & \ddots & \\ & & \sigma\lambda & (1-\sigma)\lambda \end{bmatrix}_{(i+1)\times(i+2)}, \tag{4.22}$$

$$Q_{0,0} = \begin{bmatrix} -\lambda \end{bmatrix}, \tag{4.23}$$

$$Q_{i,i} = \begin{bmatrix} -\lambda-\mu_1 & & & & & \\ & -\lambda-\mu_1-\mu_2 & & & & \\ & & \ddots & & & \\ & & & -\lambda-\mu_1-\mu_2 & & \\ & & & & -\lambda-\mu_2 \end{bmatrix}_{(i+1)\times(i+1)}, \tag{4.24}$$

$$Q_{i,i-1} = \begin{bmatrix} \mu_1 & & & \\ \mu_2 & \mu_1 & & \\ & \ddots & \ddots & \\ & & \mu_2 & \mu_1 \\ & & & \mu_2 \end{bmatrix}_{(i+2)\times(i+1)}. \tag{4.25}$$

Notice here that $Q(P,P-1)$ and $Q(P,P)$ in (4.21) are replaced by $\Lambda^P DQ_{P,P-1}$ and $\Lambda^P DQ_{P,P} + \Lambda^{P+1} DQ_{-1}$ to make the connection with the repeating equations in Section 4.4.1 (and introducing $\underline{c}$). This is found as follows: the last two boundary equations read

$$0 = \underline{\pi}_{P-2} Q_{P-2,P-1} + \underline{\pi}_{P-1} Q_{P-1,P-1} + \underline{\pi}_P Q_{P,P-1}, \tag{4.26}$$
$$0 = \underline{\pi}_{P-1} Q_{P-1,P} + \underline{\pi}_P Q_{P,P} + \underline{\pi}_{P+1} Q_{-1}, \tag{4.27}$$

with already a part of the repeating equations in the last term. After using (4.20) this becomes

$$0 = \underline{\pi}_{P-2} Q_{P-2,P-1} + \underline{\pi}_{P-1} Q_{P-1,P-1} + \underline{c}\Lambda^P DQ_{P,P-1}, \tag{4.28}$$
$$0 = \underline{\pi}_{P-1} Q_{P-1,P} + \underline{c}\Lambda^P DQ_{P,P} + \underline{c}\Lambda^{P+1} DQ_{-1}, \tag{4.29}$$

or after some rewriting

$$0 = \underline{\pi}_{P-2} Q_{P-2,P-1} + \underline{\pi}_{P-1} Q_{P-1,P-1} + \underline{c}(\Lambda^P DQ_{P,P-1}), \tag{4.30}$$
$$0 = \underline{\pi}_{P-1} Q_{P-1,P} + \underline{c}(\Lambda^P DQ_{P,P} + \Lambda^{P+1} DQ_{-1}). \tag{4.31}$$

The system of equations for solving the steady-state boundary probabilities $\tilde{\underline{\pi}} Q = 0$ with $\tilde{\underline{\pi}} = \left[\underline{\pi}_0, \cdots, \underline{\pi}_{P-1}, \underline{c}\right]$, is given by

$$0 = \underline{\pi}_0 Q_{0,0} + \underline{\pi}_1 Q_{1,0}, \tag{4.32}$$
$$0 = \underline{\pi}_{n-1} Q_{n-1,n} + \underline{\pi}_n Q_{n,n} + \underline{\pi}_{n+1} Q_{n+1,n}), n = 1, \cdots, P-2 \tag{4.33}$$
$$0 = \underline{\pi}_{P-2} Q_{P-2,P-1} + \underline{\pi}_{P-1} Q_{P-1,P-1} + \underline{c}\Lambda^P DQ_{P,P-1}), \tag{4.34}$$
$$0 = \underline{\pi}_{P-1} Q_{P-1,P} + \underline{c}(\Lambda^P DQ_{P,P} + \Lambda^{P+1} DQ_{-1}). \tag{4.35}$$

The MCF algorithm transforms this second-order vector-matrix difference equation into a first-order recurrence scheme [7]. In our case, this first-order

recurrence scheme is

$$\underline{\pi}_{n+1} = \underline{\pi}_n R_n, 0 \leq n < P-2, \tag{4.36}$$
$$\underline{c} = \underline{\pi}_{P-1} R_{P-1}. \tag{4.37}$$

Substituting the recursions into (4.32) to (4.35) yields

$$0 = \underline{\pi}_0 (Q_{0,0} + R_0 Q_{1,0}), \tag{4.38}$$
$$0 = \underline{\pi}_{n-1}(Q_{n-1,n} + R_{n-1}Q_{n,n} + R_{n-1}R_n Q_{n+1,n}), n = 1, \cdots, P-2, \tag{4.39}$$
$$0 = \underline{\pi}_{P-2}(Q_{P-2,P-1} + R_{P-2}Q_{P-1,P-1} + R_{P-2}R_{P-1}\Lambda^P D Q_{P,P-1}), \tag{4.40}$$
$$0 = \underline{\pi}_{P-1}(Q_{P-1,P} + R_{P-1}(\Lambda^P D Q_{P,P} + \Lambda^{P+1} D Q_{-1})). \tag{4.41}$$

The first equations can be used to compute R_n recursively

$$R_{P-1} = -Q_{P-1,P}(\Lambda^P D Q_{P,P} + \Lambda^{P+1} D Q_{-1})^{-1}, \tag{4.42}$$
$$R_{P-2} = -Q_{P-2,P-1}(Q_{P-1,P-1} + R_{P-1}\Lambda^P D Q_{P,P-1})^{-1}, \tag{4.43}$$
$$R_{n-1} = -Q_{n-1,n}(Q_{n,n} + R_n Q_{n+1,n})^{-1}, n = 1, \cdots, P-2. \tag{4.44}$$

The algorithm consists of first computing R_{P-1} and then calculating $R_n, n = P-2, \cdots, 1$, recursively. Using recursions (4.36) and (4.37), $\underline{\pi}_n, n = 0, \cdots, P-1$, and $\underline{c}$ can be computed recursively in terms of $\underline{\pi}_0$. In practical experiments, we will then first set $\underline{\pi}_0 = [1]$ as is often done in literature. Afterwards we have to normalize the results. Notice that we will use both the results from Sections 4.4.1 and 4.4.2 to normalize the final result. The normalization condition is given by

$$\begin{aligned}\sum_{n=0}^{\infty} \underline{\pi}_n \underline{e} &= \sum_{n=0}^{P-1} \underline{\pi}_n \underline{e} + \sum_{n=P}^{\infty} \underline{\pi}_n \underline{e} \\ &= \sum_{n=0}^{P-1} \underline{\pi}_n \underline{e} + \underline{c}\Lambda^P (I-\Lambda)^{-1} D \underline{e} = 1,\end{aligned} \tag{4.45}$$

where $\underline{e}$ is the column vector $\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}$ and I is the identity matrix. With these

results we can again calculate the mean system occupancy

$$\begin{aligned}\overline{N} &= \sum_{i=0}^{\infty} i\underline{\pi}_i\underline{e} \qquad (4.46)\\ &= \sum_{i=1}^{P-1} i\underline{\pi}_i\underline{e} + \sum_{i=P}^{\infty} i\underline{\pi}_i\underline{e}\\ &= \sum_{i=1}^{P-1} i\underline{\pi}_i\underline{e} + \sum_{i=P}^{\infty} ic\Lambda^i D\underline{e}\\ &= \sum_{i=1}^{P-1} i\underline{\pi}_i\underline{e} + c\Lambda^P (I-\Lambda)^{-2} D\underline{e} + (P-1)c\Lambda^P (I-\Lambda)^{-1} D\underline{e}\\ &= \sum_{i=1}^{P-1} i\underline{\pi}_i\underline{e} + c\Lambda^P (I-\Lambda)^{-2} D\underline{e} + (P-1)(1-\sum_{i=0}^{P-1}\pi_i\underline{e}).\end{aligned}$$

Using Little's law we can also calculate the mean delay of a customer

$$T = \frac{\overline{N}}{\lambda}. \qquad (4.47)$$

4.5 Discussion of results and numerical examples

In the remainder of this section, we discuss the results obtained in the previous sections, from a quantitative and a qualitative perspective, by means of some numerical examples. Before discussing the results, we recall the definition

$$\omega \triangleq \frac{\frac{\sigma}{\mu_1}}{\frac{\sigma}{\mu_1} + \frac{1-\sigma}{\mu_2}} = \frac{\rho_1}{\rho_1+\rho_2}. \qquad (4.48)$$

This parameter will allow us to interpret the results more intuitively; it represents the relative load of customers of type 1. This parameter can, for instance, be introduced in the stability condition (4.4), yielding

$$\rho < \frac{\left(1+\frac{1-\omega}{\omega}\right)\left(1-\left(\frac{1-\omega}{\omega}\right)^P\right)}{1-\left(\frac{1-\omega}{\omega}\right)^{P+1}}. \qquad (4.49)$$

We want to point out here that the model with a global FCFS service discipline, discussed in detail in Chapter 2, is the case where $P = 2$ (lower bound). The model without global FCFS service discipline is the case where $P = \infty$ (upper bound).

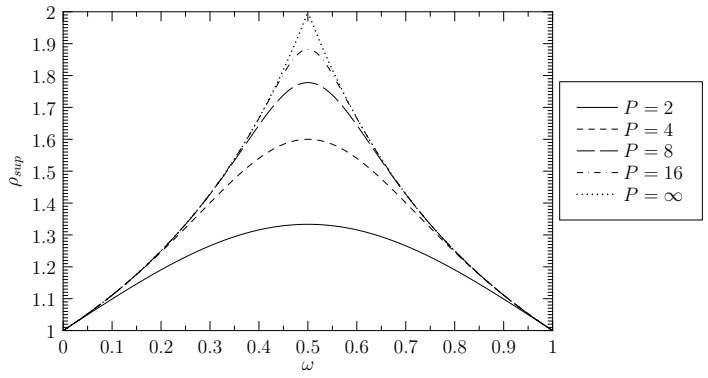


Figure 4.5: Least upper bound of the set of ρ_{sup} values where the system is stable, versus ω

In the remainder, we will first show the influence of the load (ρ) and the load balance (ω) in the system. Secondly, we will zoom in on the impact of customers of one type on customers of the other type. Finally, we will demonstrate how the result in this chapter could be used for dimensioning purposes.

Impact of load and load balance (parameters ρ and ω)

Figure 4.5 shows the influence of the load balance on the stability condition. We have plotted ρ_{sup} versus ω. Here, ρ_{sup} is the least upper bound or supremum of the set of ρ values where the system is stable and ω represents the load balance as defined in equation (4.48). The impact of parameter P is the largest when we reach the maximum for ρ_{sup} at $\omega = \frac{1}{2}$ or when the system is well balanced. A well balanced system is a system where both customers introduce the same average amount of work in the system. If the system is completely out of balance the impact of P-gFCFS becomes negligible, which is also intuitively clear since we then approach a system with almost only one type of customers and thus a single server system.

Figures 4.6 and 4.7 show the impact of load (ρ) on the mean system occupancy (for $\omega = 0.5$ and $\omega = 0.8$). Both figures show that for a small total load (ρ) the impact of P-gFCFS on the mean system occupancy becomes negligible. The impact of P-gFCFS becomes more and more noticeable when the total load increases. This is intuitively clear. In cases that the demand of the arrival stream is considerably less than what can be handled by one server, the question whether the second server is also active or not, is not very relevant. However, in cases that the demand of the arrival stream is close to or more than what can be handled by one server, the question

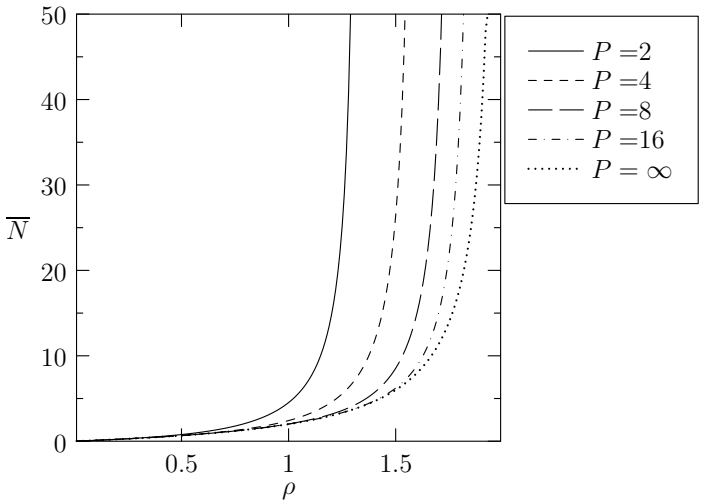


Figure 4.6: Mean system occupancy versus ρ with $\omega = 0.5$, $\mu_1 = 1$ and $\mu_2 = 4$

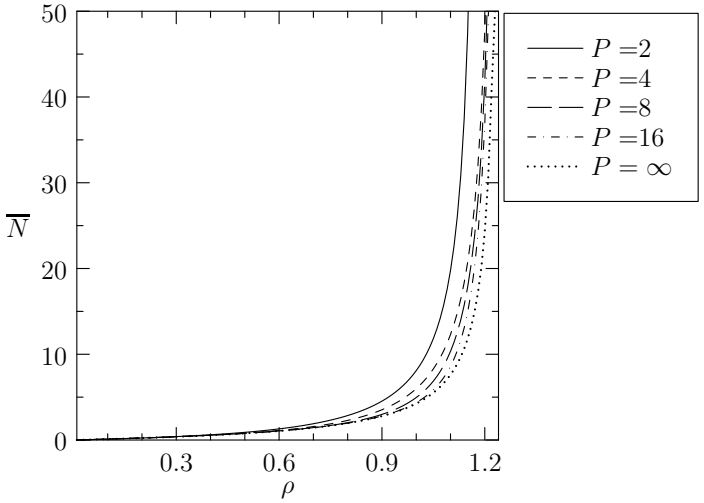


Figure 4.7: Mean system occupancy versus ρ with $\omega = 0.8$, $\mu_1 = 1$ and $\mu_2 = 4$

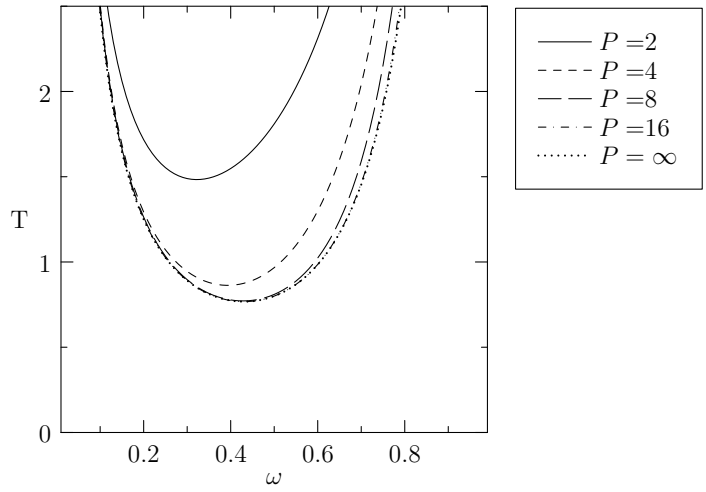


Figure 4.8: Mean customer delay versus ω with $\rho = 1$, $\mu_1 = 1$ and $\mu_2 = 4$

whether the second server is also active or not, is very relevant. In these cases the impact of P-gFCFS becomes more noticeable. Comparison of both figures also confirms the impact of load balancing (ω) on the impact of P-gFCFS. The impact of P-gFCFS is considerably larger when the system is well balanced when we consider large loads (ρ).

In Fig. 4.8 the mean customer delay versus ω with $\rho = 1$, $\mu_1 = 1$ and $\mu_2 = 4$ is shown. We see that a well balanced system ($\omega = 0.5$) no longer gives the best result when we deal with a total load (ρ) smaller than the maximum throughput. A system where the fastest server gets a higher relative load performs better than the well balanced system (see Section 2.3). When P increases, the best performing system is a more balanced system. This is again intuitively clear. When P increases, the system approaches the system without a gFCFS service discipline and the system without a gFCFS service discipline performs best when well-balanced (see Section 2.3).

Impact of customers of one type on customers of the other type

Fig. 4.9 shows the influence of the load of one type of customers on the load of the other type of customers. In both figures we have plotted ρ_2 ($= \frac{(1-\sigma)\lambda}{\mu_2}$) versus ρ_1 ($= \frac{\sigma\lambda}{\mu_1}$). The ρ_2 in both figures is the least upper bound of the set of ρ_2 values where the system is stable, for a given ρ_1 value. Here, we see that for $P = 2$, ρ_1 has a huge impact on ρ_2. This impact decreases when P becomes larger. In a road traffic context, this is exactly what we want to realise with the turn lanes. We want to decrease the impact of the vehicles

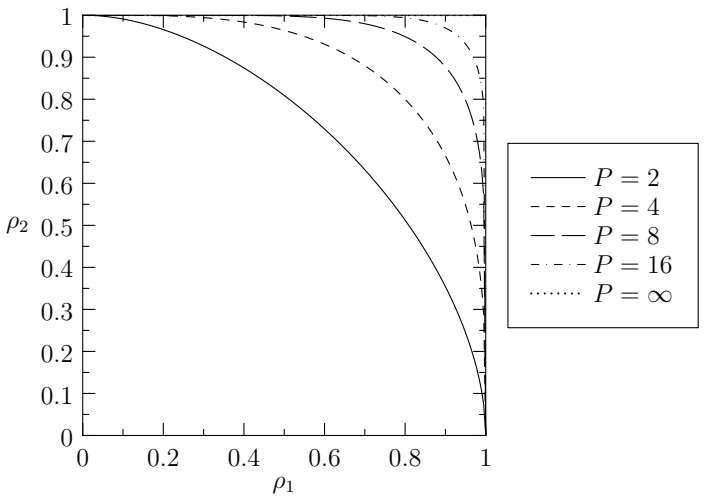


Figure 4.9: Least upper bound of the set of ρ_2 values where the system is stable, for a given ρ_1 value

going to destination 1 on vehicles with destination 2 and vice versa.

Use for dimensioning purposes

In this subsection we focus again more on the practical application of the model. Here some dimensioning possibilities are considered concerning the length of the turn lane.

This queueing system can be considered as an appropriate model for the presorting lane. We briefly explain why. First of all, we considered blocking or blockage as blocking of the server (vehicle is not able to make its turn although his destination lane is free) which is not the same as blocking of the turn lane (lane blockage or lane overflow). Secondly, there is no immediate one-to-one mapping between the model and the physical junction: the leading customers are not necessarily the vehicles on the turn lanes. When a vehicle is on the turn lane, this does not mean that the corresponding customer is necessarily a part of the leading customers. However, in case the vehicle is first in line on the turn lane, the corresponding customer *is* necessarily part of the leading customers. Therefore in both the model and physical junction, blocking only occurs when all leading customers are of the same type.

Fig. 4.10 represents the probability that at least one customer is blocked at a random time instant by customers of the other type while his own server is idle (blockage probability) versus P with $\mu_1 = 1$, $\mu_2 = 2$ and $\sigma = 0.4$.

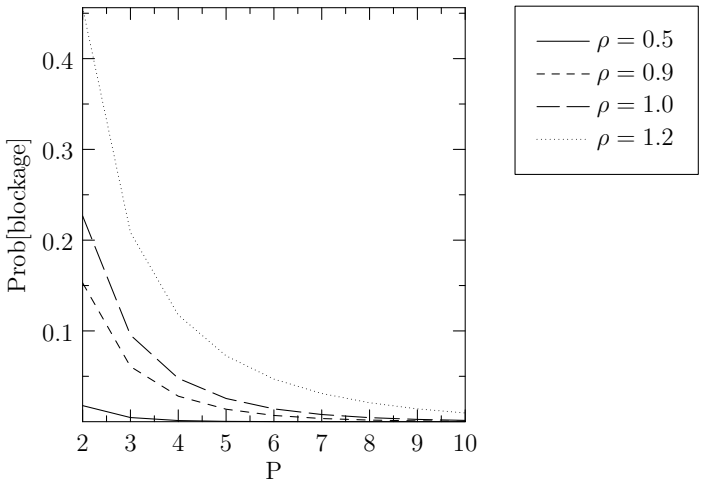


Figure 4.10: Probability that at least one customer is blocked at a random time instant while his own server is idle versus P with $\mu_1 = 1$, $\mu_2 = 2$ and $\sigma = 0.4$

In those cases, road capacity is wasted which should be avoided as much as possible. The blockage probability is given by

$$\text{Prob}\left[\text{Blockage}\right] = \sum_{n=P+1}^{\infty} \left(1 - \sigma^{n-P}\right) p(n,0) + \left(1 - (1-\sigma)^{n-P}\right) p(n,P) \tag{4.50}$$

or in words, all the probabilities where all the leading customers are of the same type multiplied with the probability that not all customers in the system are of the same type. This blockage probability represents the impact vehicles have on vehicles with another destination. This is an impact we want to reduce. As seen in Fig. 4.10, this blockage probability decreases with increasing P or increasing length of the turn lane. One possibility to determine the length of the turn lane is to determine a suitable threshold which value the blocking probability cannot exceed. If we choose, for example, the threshold value to be 0.05 then we see in Fig. 4.10 that for values of $\rho = 0.5, 0.9, 1.0, 1.2$, P should be $2, 4, 5, 7$. Notice here the similarities with the models discussed in the literature review in Section 1.4 (considering customers of type 1 are left-turning vehicles and $\mu_2 = \infty$ or through vehicles cause no delay). In those models, the blockage probability should not exceed a certain threshold value for safety reasons.

Another possibility to determine the length of the turn lane, is to satisfy a condition for the queue length, for instance, the probability that the length of the queue is longer than a certain value is at most equal to a threshold value. This is important when a traffic jam caused by the blocking effect

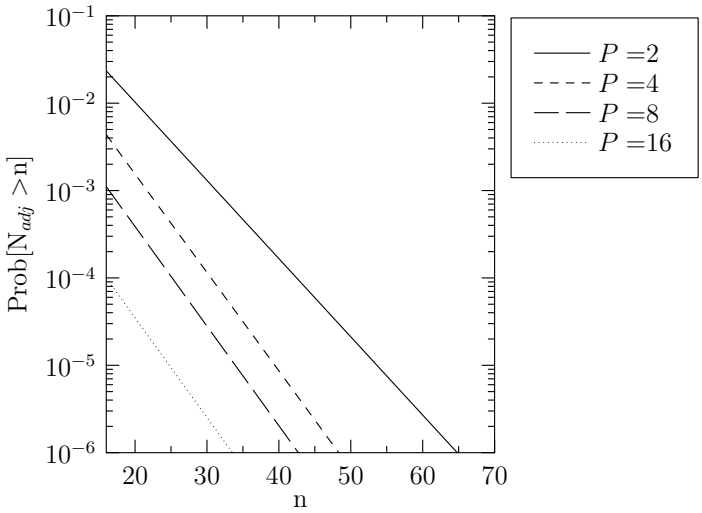


Figure 4.11: The adjusted tail probability of the system contents with $\rho = 1$, $\mu_1 = 1$, $\mu_2 = 2$ and $\sigma = 0.4$

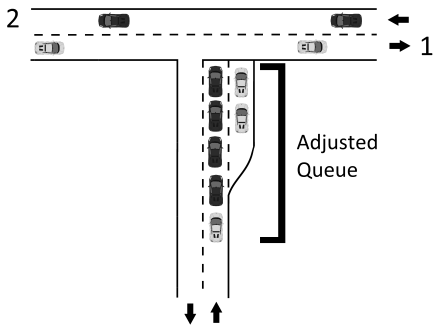


Figure 4.12: Light grey vehicles with destination 1 and dark grey vehicles with destination 2 approaching a traffic junction

can spread to other junctions, causing a domino effect. In Fig. 4.11, the adjusted tail probability of the system content with $\rho = 1$, $\mu_1 = 1$, μ_2 and $\sigma = 0.5$ is shown. The adjusted system content is the system content keeping in mind that vehicles already using the turn lane do not add to the total queue length. For example, in Fig. 4.12, the queue length is 7, but the adjusted queue length is only 5. In the example shown in Fig. 4.11, a turn lane with length 3 is required to meet the condition that the probability that the queue length is longer than 40 is not more than 10^{-5} where the pmf of the adjusted system content is defined as follows

$$\begin{aligned} p_{adj}(n) = {} & p(n,0) \\ & + \sum_{m=1}^{P-1} \left[\sum_{i=0}^{P-m-1} \left(\binom{m+i-1}{m-1} (1-\sigma)^i \sigma^m p(n+m+i,m) \right) \right. \\ & \left. + \sum_{i=0}^{m-1} \left(\binom{P-m+i-1}{P-m-1} \sigma^i (1-\sigma)^{P-m} p(n+P-m+i,m) \right) \right] \\ & + p(n,P), \end{aligned} \tag{4.51}$$

where $n \geq P$ and $\binom{m+i-1}{m-1}$ is the binomial coefficient. This formula can be understood as follows (terminology as in Fig. 4.12). The two separate lanes are blocked for further customers when the number of customers of one of the two types equals P. So if the number of customers of type 2 in the P leading customers is equal to m (and the number of customers of type 1 is $P - m$), the separate lanes are blocked from the m-th customer of type 1 or from the $P - m$-th customer of type 2 of the customers behind the leading customers, whichever comes first. The second line in (4.51) equals the probability corresponding with a blockage by a customer of type 1. This occurs, resulting in an adjusted length of n, if (i) the total number of customers equals $n + m + i$, (ii) $m - 1$ customers of the first $m + i - 1$ customers behind the leading customers are of type 1 (and $i \leq P - m - 1$ are of type 2) and (iii) the next is of type 1. (ii) and (iii) lead to a negative binomial distribution and finally line 2 of formula (4.51). The third line is due to a blockage by a customer of type 2 and can be found similarly. $p_{adj}(n)$ with $0 < n < P$ can be found analogously but is of less interest when considering the application.

The adjusted tail probability for $n \geq P$ is then given by

$$\begin{aligned}\text{Prob}[N_{adj} > n] = & \qquad (4.52)\\ & \text{Prob}[N > n]\\ & - \sum_{i=1}^{P-1} \sum_{m=1}^{P-1} \sum_{y=max(0,i-m)}^{P-m-1} \left(\frac{(y+m-1)!}{y!\,(m-1)!} (1-\sigma)^y \sigma^m p(n+i,m) \right)\\ & - \sum_{i=1}^{P-1} \sum_{m=1}^{P-1} \sum_{y=max(0,i-m)}^{m-1} \left(\frac{(y+P-m-1)!}{y!\,(P-m-1)!} (1-\sigma)^{P-m} \sigma^y p(n+i,m) \right).\end{aligned}$$

We calculated the adjusted tail probability a little bit different. When considering $\text{Prob}[N > n]$, we have to deduct some probabilities (those cases that lead to $N_{adj} \leq n$) to get $\text{Prob}[N_{adj} > n]$. We calculate these probabilities analogously as the adjusted system content.

References

[1] Leonard Kleinrock. *Queueing Systems, Volume 1: Theory.* 1975.

[2] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* 1981.

[3] Winfried Grassmann. *The Use of Eigenvalues for Finding Equilibrium Probabilities of Certain Markovian Two-Dimensional Queueing Problems.* INFORMS Journal on Computing, 15(4):412–421, 2003.

[4] James H. Wilkinson. *The Algebraic Eigenvalue Problem*, volume 87. 1965.

[5] Richard H. Gail, Sidney L. Hantler, and Alan B. Taylor. *Use of Characteristic Roots for Solving Infinite State Markov Chains.* In Computational Probability, pages 205–255. 2000.

[6] Jacques C.F. Sturm. *Mémoire sur la résolution des équations numériques.* Mém savans etrang edition, 1985.

[7] Hendrik Baumann and Werner Sandmann. *Numerical Solution of Level Dependent Quasi-Birth-and-Death Processes.* Procedia Computer Science, 1(1):1561–1569, 2010.

[8] Tuan Phung-Duc, Hiroyuki Masuyama, Shoji Kasahara, and Yutaka Takahashi. *A Simple Algorithm for the Rate Matrices of Level-Dependent QBD Processes.* In Proceedings of the 5th International Conference on Queueing Theory and Network Applications, pages 46–52, 2010.

5 Conclusions

5.1 Introduction

In the final chapter of this dissertation, the main conclusions are recapitulated in Section 5.2. Finally, the research is not finished with this dissertation. Some adjustments and enrichments can be added to the model and these are discussed in Section 5.3.

5.2 Main conclusions

In Chapter 2 we focused on the impact of the global First-Come-First-Served (gFCFS) service discipline, i.e, all arriving customers are accommodated in one single queue with dedicated servers and are served in the order of their arrival regardless of their type. We have studied two different systems where the first system uses a gFCFS service discipline and the second system adopts a FCFS service discipline for each type of customer separately. We have derived explicit closed-form formulas for the distributions of the system occupancies and customer delays for both systems (also for both types separately). The comparison of both systems allowed us to uncover the negative impact of global FCFS. Even in the case where it looked as if the impact on the maximum allowable load was negligible, we have shown the negative impact of the gFCFS service discipline on the system time of the minority of the customers. We have also shown that when the system

has to handle a high load, a well balanced system (where both customers accommodate for half of the total load) performs best. This is also apparent when investigating the stability condition. However, when the system has to handle a small total load, the system where the fastest server is more preferred, gives more performant results. This is contrary to the system without blocking where the well-balanced system always performs better. The larger the difference between the service times of the servers, the more the fastest server should be preferred.

In Chapter 3 we have shifted our focus on class clustering, i.e., customers of any given type may (or may not) have a tendency to "arrive back-to-back". A concept often neglected in literature but which we believe has a considerable impact on the performance of multiclass queueing systems. In the first part of Chapter 3, the types of customers are correlated in time according to one cluster parameter α. From a conceptual point of view, the only price we pay with this choice is that we can only study cases where both classes of customers are equiprobable and thus both types of customers account for half of the total load of the system. But due to the introduction of symmetry in the system we were able to propose a conceptual model that was still rich enough to capture the essential aspects of the problem at hand. This model allowed us to derive an explicit closed-form formula for the distributions of the system occupancy and the system delay. This allowed us to uncover the (negative) impact of the combination of global FCFS and class clustering. Class clustering is a concept that often is neglected, but we showed that it can have a considerable impact on a system and ignoring it can cause a considerable overestimation (or underestimation) of the performance. In the second part of Chapter 3, the type of the arriving customer is determined by a two-state Markov chain. We have derived an expression for the steady-state pgf of the system occupancy. We have again illustrated and quantified that it is not possible to ignore the concept of class clustering for our system. We have also shown that when we look at the stability condition or when the system has to handle a high load, a well balanced system (where both customers accommodate for half of the total load) performs not always best (only when the types of consecutive customers in the arrival stream are uncorrelated). This is especially the case when the difference in service rates of both servers is large. The system where the fastest server is more preferred, often gives more performant results. The bigger the difference between the service times of the servers, the more the fastest server should be preferred. Only when there is a negative correlation in the types of consecutive arriving customers (the correlation parameter K is smaller than 1) the slower server should be preferred (if possible).

Finally, in Chapter 4 we covered the concept of gFCFS service discipline with presorting (P-gFCFS), i.e., all arriving customers are accommodated in one single FCFS queue, regardless of their type, with an exception of the first P customers the FCFS rule holds only within the type, i.e., customers of different types can overtake each other in order to be served. We have derived an expression for the steady-state pmf of the system occupancy. We have shown that for a small total load (ρ) the impact of P-gFCFS on the mean system occupancy becomes negligible. The impact of P-gFCFS becomes more and more noticeable when the total load increases. When P increases, the best performing system is a more balanced system. Also, when P increases, the impact of one type of customer on the other type of customers decreases (which is what we want in a traffic context). In Chapter 4 we have also presented some interesting dimensioning possibilities concerning the length of the turn lane.

5.3 Further research

This dissertation presents a basic model that already provides a lot of insight into the blocking effect caused by the gFCFS scheduling. However, this model can be used as a start point for further research and can be extended in several ways to better reflect reality. One possibility is to focus more on the economic aspect and to construct meaningful soft and hard constraints to determine whether or not a turn lane is needed based on the results of this dissertation. A first possible extension is to consider general service times instead of exponential service times to better model reality. Another extension is to consider multiple dedicated servers per type to model multiple lanes. Finally, it can be worthwhile to consider multiple types. Now we consider only two types but for instance, in a road traffic context, it is possible that left turning vehicles, right turning vehicles and through vehicles need to be separated.

A

Proofs concerning zeroes in Chapter 2

In this Appendix we will prove that of the four zeroes of the form

$$\frac{a_1 \pm_s \sqrt{a_2 \pm_t 2\sqrt{a_3}}}{2\lambda}$$

with

$$\begin{aligned} a_1 &= \lambda + \mu_1 + \mu_2, \\ a_2 &= (\lambda + \mu_1 + \mu_2)^2 - 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2) - 2\mu_1\mu_2, \\ a_3 &= \mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2), \end{aligned}$$

where the signs $\pm_s$ or $\pm_t$ can be plus or minus (so four options for four zeroes), only the zero where $\pm_s = -$ and $\pm_t = +$ is inside the closed unit disk when the stability condition is met. We also prove that all zeroes are on the positive real axis.

In this Appendix we denote the zeroes by

$$\begin{array}{llllll} \hat{z}_0 & \text{where} & \pm_s = - & \text{and} & \pm_t = +, \\ \hat{z}_1 & \text{where} & \pm_s = - & \text{and} & \pm_t = -, \\ \hat{z}_2 & \text{where} & \pm_s = + & \text{and} & \pm_t = +, \\ \hat{z}_3 & \text{where} & \pm_s = + & \text{and} & \pm_t = -. \end{array}$$

The stability condition is given by

$$\lambda\left(\frac{\sigma}{\mu_1} + \frac{(1-\sigma)}{\mu_2}\right) < \frac{(\sigma\mu_2 + (1-\sigma)\mu_1)^2}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2}. \tag{A.1}$$

A.1 Proof all zeroes are real

To prove that all zeroes are real, we construct the following Sturm sequence

$$\begin{aligned}
d_0(x) &= 1, \\
d_1(x) &= -d_0(x)\frac{q_{0,0}(x)}{q_{1,0}(x)}, \\
d_2(x) &= -\frac{1}{q_{2,1}(x)}\left(d_0(x)q_{0,1}(x) + d_1(x)q_{1,1}(x)\right), \\
d_3(x) &= -\frac{1}{q_{3,2}(x)}\left(d_1(x)q_{1,2}(x) + d_2(x)q_{2,2}(x)\right),
\end{aligned}$$

where

$$\begin{aligned}
q_{i,i-1}(x) &= \sigma\mu_2 x^2, && i = 1,2,3, \\
q_{i-1,i}(x) &= (1-\sigma)\,\mu_1 x^2, && i = 1,2,3, \\
q_{0,0}(x) &= \lambda - (\lambda+\mu_1)\,x + \sigma\mu_1 x^2, \\
q_{1,1}(x) &= \lambda - (\lambda+\mu_1+\mu_2)\,x + (\sigma\mu_1 + (1-\sigma)\,\mu_2)\,x^2, \\
q_{2,2}(x) &= \lambda - (\lambda+\mu_2)\,x + (1-\sigma)\,\mu_2 x^2.
\end{aligned}$$

If we define

$$Q(x) \triangleq d_3(x)\prod_{i=0}^{2} -\sigma\mu_2 x^2,$$

notice that

$$D(z) = \lambda\frac{z^5}{1-z}Q\left(\frac{1}{z}\right),$$

where $D(z)$ is the denominator of the pgf $P(z)$ defined in equation 2.26 in Chapter 2. If x is an eigenvalue or zero of $d_3(x)$, then $z = \frac{1}{x}$ is a zero of $D(z)$. Theorem 3 in [1] states that if all zeroes of $m_i(x) = q_{i,i-1}(x)q_{i-1,i}(x) = (1-\sigma)\,\sigma\mu_1\mu_2 x^4$ are equal, then all eigenvalues are real and greater or equal to this zero. Moreover, unless this zero $a = 0$ and at least one of the $d_i(a)$ is zero, then all eigenvalues are distinct. So we can conclude that in our case all three eigenvalues are distinct and on the positive real axis. So we know that there are three real zeroes outside the closed unit disk. Consequently, there is one real zero left. In the next Section, we will prove that $\hat{z}_0$ is inside the closed unit disk. Consequently, $\hat{z}_1$, $\hat{z}_2$ and $\hat{z}_3$ are outside the closed unit disk.

A.2 Proof only $\hat{z}_0$ is inside the closed unit disk

A.2.1 Zero $\hat{z}_0$ is on the positive real axis

We prove that

$$\hat{z}_0 > 0.$$

rewriting gives us

$$\begin{aligned}
&\Leftrightarrow \quad \frac{a_1 - \sqrt{a_2 + 2\sqrt{a_3}}}{2\lambda} > 0 \\
&\Leftrightarrow \quad a_1 - \sqrt{a_2 + 2\sqrt{a_3}} > 0 \qquad\qquad (\lambda > 0) \\
&\Leftrightarrow \quad a_1 > \sqrt{a_2 + 2\sqrt{a_3}}
\end{aligned}$$

Keeping in mind that we proved in Section A.1 that $\hat{z}_0$ is real,

$$\begin{aligned}
&\Leftrightarrow \quad (a_1)^2 > \left(\sqrt{a_2 + 2\sqrt{a_3}}\right)^2 \qquad (\lambda > 0, \mu_1 > 0, \mu_2 > 0) \\
&\Leftrightarrow \quad \frac{(a_1)^2 - a_2}{2} > \sqrt{a_3}
\end{aligned}$$

and if we take into account

$$\begin{aligned}
a_1 &= \lambda + \mu_1 + \mu_2, \\
a_2 &= (\lambda + \mu_1 + \mu_2)^2 - 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2) - 2\mu_1\mu_2, \\
a_3 &= \mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2),
\end{aligned}$$

we get

$$\begin{aligned}
&\Leftrightarrow \quad \mu_1\mu_2 + 2\lambda(\sigma\mu_1 + (1-\sigma)\mu_2) > \sqrt{\mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2)} \\
&\Leftrightarrow \quad (\mu_1\mu_2 + 2\lambda(\sigma\mu_1 + (1-\sigma)\mu_2))^2 > \left(\sqrt{\mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2)}\right)^2 \\
&\qquad\qquad (\lambda > 0, \mu_1 > 0, \mu_2 > 0, 0 < \sigma < 1) \\
&\Leftrightarrow \quad \mu_1^2\mu_2^2 + 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2)\mu_1\mu_2 + (2\lambda(\sigma\mu_1 + (1-\sigma)\mu_2))^2 \\
&\qquad > 4\lambda^2\sigma(1-\sigma)\mu_1\mu_2 + \mu_1^2\mu_2^2 \\
&\Leftrightarrow \quad 4\lambda^2\left(\sigma^2\mu_1^2 + 2\sigma(1-\sigma)\mu_1\mu_2 + (1-\sigma)^2\mu_2^2\right) + \mu_1^2\mu_2^2 \\
&\quad + 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2)\mu_1\mu_2 > 4\lambda^2\sigma(1-\sigma)\mu_1\mu_2 + \mu_1^2\mu_2^2 \\
&\Leftrightarrow \quad 4\lambda^2\left(\sigma^2\mu_1^2 + \sigma(1-\sigma)\mu_1\mu_2 + (1-\sigma)^2\mu_2^2\right) \\
&\quad + 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2)\mu_1\mu_2 > 0
\end{aligned}$$

which concludes our proof since $\lambda > 0$, $\mu_1 > 0$, $\mu_2 > 0$ and $0 < \sigma < 1$.

A.2.2 Lemma

Here we prove that

$$2\lambda - a_1 < 0$$

or if we take into account that

$$a_1 = \lambda + \mu_1 + \mu_2.$$

we get

$$\lambda - \mu_1 - \mu_2 < 0,$$

or

$$\lambda < \mu_1 + \mu_2.$$

First notice that the stability condition (A.1) can be written as

$$\lambda < \frac{\mu_1\mu_2(\sigma\mu_2 + (1-\sigma)\mu_1)}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2}.$$

Thus by proving

$$\frac{\mu_1\mu_2(\sigma\mu_2 + (1-\sigma)\mu_1)}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2} < \mu_1 + \mu_2,$$

it follows that

$$2\lambda - a_1 < 0.$$

This proof is as follows

$$\begin{aligned}
\Leftrightarrow \quad & \frac{\mu_1\mu_2(\sigma\mu_2 + (1-\sigma)\mu_1)}{(\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2} < \mu_1 + \mu_2 \\
\Leftrightarrow \quad & \mu_1\mu_2(\sigma\mu_2 + (1-\sigma)\mu_1) < \\
& (\mu_1 + \mu_2)\left((\sigma\mu_2 + (1-\sigma)\mu_1)^2 - \sigma(1-\sigma)\mu_1\mu_2\right) \\
\Leftrightarrow \quad & \sigma\mu_1\mu_2^2 + (1-\sigma)\mu_1^2\mu_2 < (\sigma\mu_2 + (1-\sigma)\mu_1)^2\mu_1 \\
& \quad - \sigma(1-\sigma)\mu_1^2\mu_2 + (\sigma\mu_2 + (1-\sigma)\mu_1)^2\mu_2 - \sigma(1-\sigma)\mu_1\mu_2^2 \\
\Leftrightarrow \quad & \sigma\mu_1\mu_2^2 + (1-\sigma)\mu_1^2\mu_2 < \sigma^2\mu_1\mu_2^2 + \sigma(1-\sigma)\mu_1^2\mu_2 \\
& \quad + (1-\sigma)^2\mu_1^3 + \sigma^2\mu_2^3 + \sigma(1-\sigma)\mu_1\mu_2^2 + (1-\sigma)^2\mu_1^2\mu_2 \\
\Leftrightarrow \quad & 0 < (1-\sigma)^2\mu_1^3 + \sigma^2\mu_2^3 + \mu_1\mu_2^2(-\sigma + \sigma^2 + \sigma(1-\sigma)) \\
& \quad + \mu_1^2\mu_2(-(1-\sigma) + (1-\sigma)^2 + \sigma(1-\sigma)) \\
\Leftrightarrow \quad & 0 < (1-\sigma)^2\mu_1^3 + \sigma^2\mu_2^3
\end{aligned}$$

which concludes our intermezzo since $\mu_1 > 0$, $\mu_2 > 0$ and $0 < \sigma < 1$.

A.2.3 Zero $\hat{z}_0$ is inside the closed unit disk when the stability condition is met

We need to prove that

$$|\hat{z}_0| < 1.$$

Since $\hat{z}_0$ is positive, we can write

$$\Leftrightarrow \quad \hat{z}_0 < 1 \qquad \texttt{(A.2.1)}$$

Rewriting gives us

$$\begin{aligned}
&\Leftrightarrow \quad \frac{a_1 - \sqrt{a_2 + 2\sqrt{a_3}}}{2\lambda} < 1 \\
&\Leftrightarrow \quad a_1 - \sqrt{a_2 + 2\sqrt{a_3}} < 2\lambda \qquad (\lambda > 0) \\
&\Leftrightarrow \quad -\sqrt{a_2 + 2\sqrt{a_3}} < 2\lambda - a_1 \\
&\Leftrightarrow \quad \left(-\sqrt{a_2 + 2\sqrt{a_3}}\right)^2 > (2\lambda - a_1)^2 \qquad \texttt{(A.2.2)} \\
&\Leftrightarrow \quad \sqrt{a_3} > \frac{(2\lambda - a_1)^2 - a_2}{2}
\end{aligned}$$

and if we take into account

$$\begin{aligned}
(2\lambda - a_1)^2 &= (\lambda - \mu_1 - \mu_2)^2 \\
&= (\lambda + \mu_1 + \mu_2)^2 - 4\lambda(\mu_1 + \mu_2), \\
a_2 &= (\lambda + \mu_1 + \mu_2)^2 - 4\lambda(\sigma\mu_1 + (1-\sigma)\mu_2) - 2\mu_1\mu_2, \\
a_3 &= \mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2),
\end{aligned}$$

we get

$$\Leftrightarrow \quad \sqrt{\mu_1\mu_2(4\lambda^2\sigma(1-\sigma) + \mu_1\mu_2)} > \mu_1\mu_2 - 2\lambda((1-\sigma)\mu_1 + \sigma\mu_2)$$

At this point we have to split up our proof, namely in a part where $\mu_1\mu_2 - 2\lambda((1-\sigma)\mu_1 + \sigma\mu_2) < 0$ and $\mu_1\mu_2 - 2\lambda((1-\sigma)\mu_1 + \sigma\mu_2) > 0$. In the first part where $\mu_1\mu_2 - 2\lambda((1-\sigma)\mu_1 + \sigma\mu_2) < 0$ we can conclude our proof. However if $\mu_1\mu_2 - 2\lambda((1-\sigma)\mu_1 + \sigma\mu_2) > 0$, we can take the square of both

sides of the inequality. We get

$$
\begin{aligned}
&\Leftrightarrow \quad \left(\sqrt{\mu_1\mu_2(4\lambda^2\sigma(1-\sigma)+\mu_1\mu_2)}\right)^2 > (\mu_1\mu_2 - 2\lambda((1-\sigma)\mu_1+\sigma\mu_2))^2 \\
&\Leftrightarrow \quad 4\lambda^2\sigma(1-\sigma)\mu_1\mu_2 + \mu_1^2\mu_2^2 \\
&\qquad > \mu_1^2\mu_2^2 - 4\lambda((1-\sigma)\mu_1+\sigma\mu_2)\mu_1\mu_2 + 4\lambda^2((1-\sigma)\mu_1+\sigma\mu_2)^2 \\
&\Leftrightarrow \quad 4\lambda((1-\sigma)\mu_1+\sigma\mu_2)\mu_1\mu_2 - 4\lambda^2((1-\sigma)\mu_1+\sigma\mu_2)^2 \\
&\qquad + 4\lambda^2\sigma(1-\sigma)\mu_1\mu_2 > 0 \\
&\Leftrightarrow \quad \lambda < \frac{((1-\sigma)\mu_1+\sigma\mu_2)\mu_1\mu_2}{((1-\sigma)\mu_1+\sigma\mu_2)^2 - (\sigma(1-\sigma)\mu_1\mu_2)} \\
&\Leftrightarrow \quad \lambda\left(\frac{\sigma}{\mu_1} + \frac{(1-\sigma)}{\mu_2}\right) < \frac{(\sigma\mu_2+(1-\sigma)\mu_1)^2}{(\sigma\mu_2+(1-\sigma)\mu_1)^2 - (\sigma(1-\sigma)\mu_1\mu_2)}
\end{aligned}
$$

which is the stability condition, which we know is met. This concludes the proof.

References

[1] Grassmann Winfried. *Real Eigenvalues of Certain Tridiagonaal Matrix Polynomials, with Queueing Applications.* Linear Algebra and its Applications, 342:93–106, 2002.