# Numerical Analysis of Multidimensional Queueing Systems 

Numerieke analyse van multidimensionale wachtlijnsystemen

## Ekaterina Evdokimova

Promotoren: prof. dr. ir. S. Wittevrongel, prof. dr. ing. D. Fiems Proefschrift ingediend tot het behalen van de graad van Doctor in de ingenieurswetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2017-2018

ISBN 978-94-6355-059-8

Vakgroep Telecommunicatie en Informatieverwerking Voorzitter: prof. dr. ir. H. Bruneel Faculteit Ingenieurswetenschappen en Architectuur Academiejaar 2017-2018

Leden van de examencommissie:

Prof. Patrick De Baets (voorzitter)
Prof. Sabine Wittevrongel (promotor)
Prof. Dieter Fiems (promotor)
Prof. Joris Walraevens
Prof. Stijn De Vuyst
Prof. Antoon Bronselaer
Dr. Balakrishna Prabhu
Prof. Alexey Vinel

## Acknowledgement

First, I would like to express my gratitude to my promotors Dieter Fiems and Sabine Wittevrongel for their guidance, their willingness to share their knowledge, and their patience. I deeply appreciate the time they dedicated to my work, ideas, and funding to make my Ph.D. experience productive. Special mention goes to Koen De Turck, I appreciate all his help and input to this dissertation. My sincere thanks also goes to Alexey Vinel, who has recommended me for this Ph.D. position here, and later became part of a fruitful research collaboration which resulted in one of the chapters of this dissertation. Moreover, I would like to thank the colleagues of SMACS group and the head of the department Herwig Bruneel, it was a great pleasure to work here. All colleagues with whom we shared office 1.18: Marijn, Apoorv, Veronique, Sofian, Patrick, Marco, Tijs, Alessandro: thank you for your company and warm atmosphere. I would like to thank Eline, a friend of mine and a former Ph.D. student of SMACS, who was the first person to meet me and show me around after I arrived in Belgium for the first time. I would also like to mention several other PhD-fellows from the neighboring departments who became my friends: Maryna, Celina, Mima, Kang, Mykola, Marina, Abhi and Pankaj. Thank you guys, for all the fun we had together and all the trouble we shared. I went to Maryna and Erland's to exercise my presentations, so I am grateful for their feedback and general input. Moreover, I am thankful to Maryna for motivating and inspiring me and giving me a great example of how one should keep going and never give up. I am grateful for the support and help from my friends Claire, Leen and Laura. Thank you, guys, for being in my life and for all vegan dinners we cooked for each other. I also want to thank my friends in Russia Natasha, Tanya, Ilja and Ira, for bound that stays between us, even though we live so far apart.

I am also thankful for the support of my family. First of all my parents; Tonya and Oleg, who are the best parents one could possibly have. I would like to mention my grandparents: Nadejda, Evgeniya and my grandfather Yuri, who always wanted me to receive as many academic degrees as possible. I would like to bring up the memory of my grandfather Valeri and grandaunt Larisa, who passed away a few years ago, but would be happy to find their names here.

Finally, I would like to thank my girlfriend Hanne, whose faithful support in the last stages of this Ph.D. can hardly be overestimated. I am grateful for all the strength, love and understanding that I got from you when I needed it the most.

Gent, 2017
Ekaterina Evdokimova

## Table of Contents

Acknowledgement ..... i
Nederlandse samenvatting ..... $\mathbf{v}$
English summary ..... ix
1 Introduction ..... 1-1
1.1 Introduction ..... 1-1
1.2 Applications of multidimensionalqueueing models1-3
1.2.1 Opportunistic scheduling in wireless networks ..... 1-4
1.2.2 Drive-thru Internet in vehicular communication ..... 1-6
1.2.3 Assembly systems ..... 1-7
1.3 Methodology ..... 1-9
1.3.1 Markov processes. ..... 1-9
1.3.2 Numerical methods for solving queueing models ..... 1-11
1.3.2.1 Product-form solutions ..... 1-11
1.3.2.2 Matrix-geometric method ..... 1-15
1.3.2.3 Iterative methods ..... 1-17
1.3.2.4 Series expansion techniques ..... 1-20
1.4 Multidimensional queueing models. ..... 1-23
1.5 Dissertation outline ..... 1-26
1.6 Publications ..... 1-27
1.6.1 Publications in international journals ..... 1-27
1.6.2 Papers in proceedings of international conferences ..... 1-27
1.6.3 Abstracts and other presentations ..... 1-28
References ..... 1-29
2 A Taylor Series Approach for Service-Coupled Queueing Systems with ..... $-$
Intermediate Load ..... 2-1
2.1 Introduction ..... 2-2
2.2 Performance analysis ..... 2-4
2.3 Numerical results ..... 2-9
2.4 Conclusions ..... 2-15
References. ..... 2-15
3 Coupled queues with customer impatience ..... 3-1
3.1 Introduction ..... 3-2
3.2 Light traffic analysis ..... 3-4
3.2.1 Numerical series expansion ..... 3-6
3.2.2 Closed form expressions for the symmetric coupled queue- ing system ..... 3-8
3.2.3 Lower bound for the radius of convergence ..... 3-8
3.3 Overload analysis ..... 3-11
3.3.1 Numerical series expansion ..... 3-12
3.3.2 Closed form expressions for the symmetric coupled queue- ing system ..... 3-13
3.4 Fluid limit ..... 3-13
3.5 Numerical results and discussion ..... 3-16
3.6 Conclusions ..... 3-20
References. ..... 3-21
4 Queueing Analysis of Opportunistic Scheduling with Spatially Corre- lated Channels ..... 4-1
4.1 Introduction ..... 4-2
4.2 Queueing model ..... 4-5
4.3 Performance analysis ..... 4-6
4.3.1 Balance equations ..... 4-7
4.3.2 Regular perturbation ..... 4-7
4.3.3 Overload-traffic analysis ..... 4-9
4.3.4 $\quad$ Light-traffic analysis ..... 4-10
4.3.5 Performance metrics ..... 4-11
4.4 Modelling the wireless environment ..... 4-12
4.4.1 Single-channel model ..... 4-12
4.4.2 Multiple channels ..... 4-13
4.4.3 Channel State Information ..... 4-16
4.4.4 Examples of schedulers ..... 4-16
4.5 Results ..... 4-18
4.5.1 Independent channels ..... 4-18
4.5.2 Correlated channels. ..... 4-22
4.6 Conclusions ..... 4-28
References ..... 4-29
5 Internet Provisioning in VANETs: Performance Modelling of Drive-Thru Scenarios ..... 5-1
5.1 Introduction ..... 5-2
5.2 Markov model ..... 5-5
5.2.1 Modelling assumptions ..... 5-5
5.2.2 Balance equations ..... 5-7
5.2.3 Discussion ..... 5-8
5.3 Performance analysis ..... 5-10
5.3.1 A few vehicles with high load ..... 5-11
5.3.2 Perturbation of the arrival rate: $\lambda \approx \lambda_{0}$ ..... 5-12
5.3.3 Perturbation of the transmission rate $\mu \approx \mu_{0}$ ..... 5-14
5.3.4 Performance metrics ..... 5-15
5.3.4.1 Moments of the number of vehicles, and the ve- hicle blocking probability ..... 5-15
5.3.4.2 Mean queue size at the RSU ..... 5-16
5.3.4.3 Packet loss ..... 5-16
5.3.4.4 Packet delay ..... 5-17
5.4 Numerical results ..... 5-18
5.5 Conclusions ..... 5-24
References ..... 5-25
6 Conclusions ..... 6-1
6.1 Overview of the main results ..... 6-1
6.2 Future work ..... 6-2

# Nederlandse samenvatting -Summary in Dutch- 

Eenvoudig gezegd komt elke situatie waarbij "klanten" wachten in rijen om een of andere vorm van bediening te ontvangen, neer op een wachtlijnsysteem. Zodra deze klanten bediend zijn, verlaten ze ofwel het systeem of treden ze toe tot een andere wachtlijn. Zo maken ze opnieuw deel uit van een wachtlijnsysteem. De term "klant" is hier een natuurlijke abstractie, die zowel kan verwijzen naar wachtende mensen bij een kassa, naar pakjes die in het postkantoor wachten om verdeeld te worden, als naar een digitaal informatiepakket dat wacht op transmissie over een of ander transmissiekanaal, enz. Hetzelfde principe geldt voor de "bediening" die door deze klanten gevraagd wordt, die in feite een abstractie vormt voor de tijdsduur dat de klant de server bezet houdt. Zo bijvoorbeeld is het de tijd die het inscannen van de boodschappen aan de kassa in beslag neemt, de uren of minuten tot een postpakje bij het postkantoor zijn weg vindt naar de juiste bestelwagen, of de transmissietijd van een digitaal pakket.

De meeste wachtlijnsystemen delen de eigenschap dat zowel de aankomsttijdstippen van klanten als de hoeveelheid bediening die de klanten vereisen, niet op voorhand gekend zijn. Indien dit wel het geval was, zou de serviceprovider een planning voor de klanten kunnen opstellen of de noodzakelijke bedieningscapaciteit kunnen voorzien, zodat geen enkele klant nog zou hoeven te wachten. Meer concreet kunnen we het volgende voorbeeld geven: als een arts op voorhand zou weten hoeveel tijd iedere consultatie in beslag zou nemen, zouden wachttijden voor patiënten volledig vermeden kunnen worden. Iedere patiënt zou immers zijn afspraak krijgen op het moment dat de consultatie van de vorige patiënt afloopt.

Wiskundig wordt het gebrek aan kennis of de onzekerheid over aankomsttijdstippen en bedieningstijden het handigst beschreven door middel van toevalsgrootheden. Wachtlijntheorie bestudeert daarom probabilistische modellen van wachtlijnsystemen. Door het modelleren van wachtlijnsystemen kan men inzicht verwerven in de dynamica van het wachtlijnsysteem door het berekenen van diverse prestatiematen, zoals gemiddelde wachttijden en wachtlijnbezettingen. Een dergelijke prestatie-analyse is essentieel bij het ontwerp van een wachtlijnsysteem het ontwerp omvat o.a. hoeveel wachtruimte en hoeveel servers men dient te voorzien - aangezien het toelaat de verwachte prestatie in te schatten vooraleer het wachtlijnsysteem wordt gebouwd. De toevalsprocessen die de tijdsdynamica van een wachtlijnsysteem beschrijven, behoren veelal tot de klasse van de Markovprocessen. De klasse van de Markovprocessen is niet alleen een veelzijdige klasse
van stochastische processen die een breed gamma aan toevalsfenomenen kan modelleren, maar is tevens een klasse van processen waarvoor men verschillende karakteristieken daadwerkelijk kan berekenen. De voornaamste karakteristiek van een Markovproces is zijn regimedistributie. Voor Markovprocessen die wachtlijnsystemen voorstellen zoals in dit werk kunnen de meeste prestatiematen van het wachtlijnsysteem uitgedrukt worden in termen van deze regimedistributie. In wachtlijnterminologie verwijst het oplossen van een wachtlijnmodel derhalve gewoonlijk naar het vinden van de regimedistributie van het onderliggende Markovproces.

Dit proefschrift bestudeert een specifiek type van wachtlijnmodellen, de zogenaamde multidimensionele wachtlijnmodellen. Een multidimensioneel wachtlijnmodel is een model dat bestaat uit meerdere wachtlijnen met gedeelde servers. In tegenstelling tot klassieke wachtlijnmodellen met één enkele wachtlijn of netwerken van wachtlijnen met toegewezen servers waarbij de bedieningssnelheid van elke server enkel afhangt van de toestand van zijn wachtlijn, zullen we ons in dit proefschrift concentreren op modellen waarbij de inhoud van elke individuele wachtlijn een effect heeft op de prestatie van de andere wachtlijnen en van het systeem in het algemeen. Modellen met wachtlijnen die onderling afhankelijk zijn van elkaar vertonen een specifieke dynamica, die niet teruggevonden wordt bij traditionele wachtlijnsystemen. Deze modellen laten toe complexe systemen voor te stellen waarbij de interactie tussen de wachtlijnen kan veroorzaakt worden door meerdere factoren zoals blokkering of gedeelde servers. Specifiek worden in dit proefschrift vier verschillende wachtlijnmodellen voorgesteld en geanalyseerd, komende uit toepassingsscenario's in de domeinen van industriële productiesystemen en draadloze communicatie.

Het numeriek oplossen van multidimensionele wachtlijnmodellen is computationeel moeilijk, omwille van de enorme toestandsruimte van Markovprocessen met meerdere dimensies. Dit staat bekend als het probleem van toestandsruimteexplosie, wat betekent dat de toestandsruimte exponentieel groeit met het aantal betrokken dimensies (of wachtlijnen). Reeds bij enkele wachtlijnen van een matige capaciteit omvat de toestandsruimte duizenden toestanden. Een dergelijk model is dan ook moeilijk te analyseren, aangezien het betekent dat zeer grote stelsels van lineaire vergelijkingen moeten worden opgelost. Bestaande wachtlijntheoretische methoden laten toe complexe modellen op een efficiënte manier op te lossen. Deze methoden zijn echter enkel toepasbaar op Markovmodellen met een zekere vereenvoudigde structuur, die de modellen met onderling interagerende wachtlijnen typisch niet bezitten. Niettemin laten verschillende eigenschappen van de generatormatrix een vereenvoudigde oplossing toe. Bijvoorbeeld, voor alle systemen die in dit proefschrift bestudeerd worden, zijn de generatormatrices ijl omwille van de beperkte set van transities die bereikbaar zijn vanuit elk van de toestanden van het systeem. Aan de andere kant, zelfs wanneer we gebruik maken van structurele eigenschappen, stuiten we vaak op een onhandelbare computationele complexiteit van de exacte oplossing. Daarom zullen we in dit proefschrift vooral een numerieke benaderingstechniek onderzoeken die steunt op een perturbatie-analyse van de regimedistributie van het Markovproces.

Dit proefschrift bestaat uit vijf hoofdstukken. In het eerste hoofdstuk introduceren we de praktische toepassingen die aanleiding geven tot de onderzochte multidimensionele wachtlijnmodellen, evenals de wiskundige basistechnieken die in dit proefschrift gebruikt worden om deze wachtlijnmodellen op te lossen. Verdere hoofdstukken specifiëren de onderstellingen omtrent de beschouwde wachtlijnsystemen en gaan dieper in op de numerieke methoden die toegepast worden om benaderende oplossingen te bekomen. Hoewel in alle gevallen reeksontwikkelingstechnieken gebruikt worden, vraagt elk wachtlijnmodel toch een individuele aanpak om de gewenste prestatiematen te berekenen. In hoofdstuk 2 introduceren we een model voor een voorraadsysteem met meerdere producten, dat zich bevindt vóór een assemblagelijn. De voorraden aan verschillende onderdelen worden gemodelleerd als wachtlijnen, waarbij de onderdelen de rol van de klanten vervullen, terwijl de assemblagelijn zelf de server vormt van het wachtlijnsysteem. Hoofdstuk 3 bestudeert een gelijkaardig model. Als een uitbreiding van het model uit hoofdstuk 2 introduceert hoofdstuk 3 echter de bijkomende onderstelling dat alle onderdelen uit de voorraad moeten gebruikt worden vóór een zekere vervaldag. Onderdelen die niet gebruikt worden vóór dat tijdstip, worden uit de voorraad verwijderd.

Hoewel de vervaldag het enige verschilpunt is tussen de modellen in hoofdstuk 2 en hoofdstuk 3, wordt hierbij een geheel andere dynamica geobserveerd. Daar waar de dynamica in hoofdstuk 2 bepaald wordt door serverblokkering (de server blokkeert wanneer eén van de voorraden leeg is), domineert het verwijderingsproces dikwijls de dynamica in hoofdstuk 3. De laatste twee hoofdstukken behelzen wachtlijnmodellen voor een aantal specifieke scenario's bij draadloze communicatie. Hoofdstuk 4 betreft transmissies vanaf een draadloos toegangspunt naar meerdere gebruikers toe, waarbij een opportunistische scheduleringsstrategie gebruikt wordt. Het model brengt de toevallige kanaalvariaties in rekening, terwijl de prestatie-analyse de efficiëntie van verschillende scheduleringsschema's evalueert en vergelijkt. In hoofdstuk 5 richten we ons op een zogenaamd Drive-thruInternetscenario, waarbij een toegangspunt dat zich langs een weg bevindt, internetconnectie voorziet voor de mobiele gebruikers in de auto's in een omgeving met meerdere voertuigen.

Voor al deze systemen hebben we aangetoond dat numerieke reeksontwikkelingstechnieken toelaten de prestatie zowel snel als nauwkeurig te evalueren. De methodologie die hier werd ontwikkeld, is niet alleen toepasbaar op de specifieke wachtlijnsystemen die in dit proefschrift werden onderzocht, maar wordt verwacht ook toepasbaar te zijn voor een breed gamma aan wachtlijnmodellen met meerdere interagerende wachtlijnen.

## English summary

Any real-world situation which involves customers that join waiting lines in order to receive some service is a queueing system. Once the customers have received the requested service, they either leave or join another waiting line in the queueing system. The term "customer" is an abstraction which can equally refer to people waiting at the checkout counter, parcels at the post office waiting to be distributed, a digital packet with information waiting to be transmitted over some transmission channel, etc. Likewise, the"service" that is requested by these customers is an abstraction for the duration that the customer occupies the server. For example, it is the time to process the groceries at the checkout counter, the time needed to route a parcel to the correct delivery van at the post office, or the transmission time of the digital packet.

Most queueing systems share the property that arrival instants of customers as well as the amount of service they require are not known in advance. If this was the case, the service provider could either schedule the customers such that no customer will have to wait, or provide the necessary service capacity to avoid queueing. For example, if a physician would know in advance how long each patient's consultation would take, patient waiting times could be completely avoided by scheduling the next patient at the end of the consultation of the preceding patient.

Mathematically, the lack of knowledge or uncertainty about arrival times and service times is most conveniently described by random variables. Queueing theory therefore studies probabilistic models of queueing systems. By modelling queueing systems, one can gain an understanding of the queueing systems dynamics by calculating various performance measures, like expected waiting times and queue lengths. Such a performance analysis is an essential tool while designing the queueing system - the design comprises a.o. how much space one should allocate for queueing and how many servers one should provide - as it allows for estimating the expected performance prior to building the queueing system. The random processes which represent the queueing dynamics over time belong most often to the class of Markov processes. The class of Markov processes is not only a versatile class of stochastic processes which can model a wide range of random phenomena, it is also a class of processes for which one can actually calculate various characteristics. The most important characteristic of a Markov process is its stationary distribution. For Markov processes representing queueing systems like in this work, most performance measures of the queueing system can be expressed in terms of this stationary distribution. In queueing parlance, solving
a queueing model therefore usually refers to finding the stationary distribution of the underlying Markov process.

This dissertation studies a specific type of queueing models, referred to as multidimensional queueing models. A multidimensional queueing model is a model which involves multiple queues with shared servers. In contrast to classic queueing models with a single queue or networks of queues with dedicated servers where the service speed of every server only depends on the state of its queue, we focus on models where the content of each individual queue affects the performance of the other queues and the system in general. Models with dependent queues exhibit specific queueing dynamics, which are not observed in traditional queueing systems. These models allow for representing complex systems where the interaction between the queues can be caused by multiple factors like blocking or server sharing. In particular, four different queueing models are proposed and analysed in this dissertation coming from application scenarios in the areas of industrial manufacturing systems and wireless communications.

Numerically solving multidimensional queueing models is computationally hard due to the size of the state space of Markov processes with multiple dimensions. This is known as the state space explosion problem. The state space grows exponentially in the number of dimensions (or queues) involved. Already with a few queues with a modest capacity, the state space attains thousands of states. Such a model is hard to analyse, as it implies solving very large systems of linear equations. Existing methods of queueing theory allow for solving large models efficiently, however, they are applicable only to Markov models with a certain simplified structure, that the models with interacting queues typically do not have. Nevertheless, several properties of the transition rate matrix allow for a simplified solution. For example, for all the systems studied in this dissertation, the generator matrices possess sparsity due to the limited set of transitions reachable from each state of the system. On the other hand, even utilising structural properties, we often encounter unsustainable computational complexity of the exact solution. Therefore, in this dissertation we mainly investigate a numerical approximation technique which relies on a perturbation analysis of the stationary distribution of the Markov process.

This dissertation consists of five chapters. The first chapter introduces the practical applications leading to the multidimensional queueing models under investigation, as well as the basic mathematical techniques which are used in this dissertation to solve these queueing systems. Further chapters specify the assumptions regarding the queueing systems of interest and elaborate on the numerical methods applied to obtain the approximate solutions. Even though in all cases series expansion techniques are employed, each queueing model requires an individual approach to calculate the performance measures of interest. In Chapter 2 we introduce a model of a multi-product inventory located in front of an assembly line. The inventories with the parts are modelled as queues, the parts being the customers, while the assembly line itself is the server of the queueing system. Chapter 3 studies a similar model. However, as an extension of the model of Chapter 2, Chapter 3 introduces the additional assumption that all parts in the inventory
have to be used before a due date. Parts that are not used before their due date are removed from the inventory. While the expiration date is the only difference between the models in Chapter 2 and Chapter 3, completely different dynamics are observed. While in Chapter 2 the dynamics are governed by service blocking (the server blocks if one of the inventories is empty), the abandonment process often dominates the dynamics in Chapter 3. The last two chapters deal with queueing models for some specific scenarios in wireless communications. Chapter 4 focuses on multi-user downlink transmissions by a wireless access point, working under an opportunistic scheduling policy. The model accounts for the random channel variations while the performance study evaluates efficiency and compares various scheduling schemes. In Chapter 5 we address a Drive-thru Internet scenario, where an access point located along a road provides Internet connection for the mobile users in the cars in a multi-vehicular environment.

For all these systems, we have shown that numerical series expansion techniques allow for evaluating the performance both fast and accurately. The methodology developed here not only applies to the specific queueing systems that were investigated here, but is expected to apply to a wide range of queueing models involving multiple interacting queues.


## Introduction

### 1.1 Introduction

Waiting in line or queueing is not only one of the more undesirable situations in our life to be in, it is also the subject of a vast research area named queueing theory. The need for waiting in line is intuitively clear. A number of people would like the same service at the same time and arrive at the same place to find it. In old times this would immediately result in confrontation, but in modern society we respectfully wait our turn and form a structured queue or, in other words, a queueing system. Any queueing system is characterised by arrivals and departures of customers, wanting to receive some kind of service and possibly having to wait for it.

To some extent, most of us practice queueing analysis on a daily basis. Whenever we have to choose between joining different queues, we make an educated guess in which queue we have to wait the least. For example, in a supermarket we may simply opt for the waiting line with the least number of customers, or make a more refined estimation by accounting for the number of items in the waiting customers' baskets. Also while waiting in line, we perform some sort of queueing analysis. We can comfort ourselves by measuring the progress we make in the queueing system, which is equivalent to measuring the time it will take till it is our turn. For example, in a traffic jam we count down the distance to the junction or the narrowing of the road which causes the jam. Sometimes, we even compare the service speed of queues, for example by evaluating if it is worth changing queues.

The arguments one intuitively uses to estimate waiting times are made precise by queueing theory. For such a rigorous analysis, queueing theory starts with a well-defined queueing model. A queueing model clearly specifies all assumptions on customer arrivals, their service requirements and the order in which the queueing system's server attends the customers, the common assumption being that customers arrive in accordance with some well-specified stochastic process, have random service requirements and are served in order of arrival. Given these assumptions, the model can be solved, either analytically or numerically. Here solving means that we can express various performance measures of the queueing system in terms of the parameters of the arrival and service processes. An analytical solution provides a closed-form formula for these performance measures whereas a numerical solution proposes an algorithm which can efficiently calculate these performance measures. The performance measures of interest usually describe the long-term queueing behaviour and include throughput (the average number of customers that the queueing system can serve per time unit), the mean waiting time or the waiting time distribution of an arbitrary customer, the mean and the distribution of the number of waiting customers at some point in time, etc. These performance measures can then be used to dimension queueing systems. Dimensioning problems include questions like how many servers do we need to guarantee that customers do not have to wait overly long, or how much space do we need for waiting customers. In addition, queueing analysis can also help while designing a queueing system. Design questions include the placement of queues (e.g. one long queue vs. multiple dedicated queues) as well as the choice of queueing disciple (the order in which customers are served).

Since the concept of queueing analysis was introduced by A.K. Erlang in 1909 [1], queueing models found multiple applications in various fields, the most important fields being telecommunication systems and production systems. Since its inception, queueing theory has grown into a vast research field. While many important queueing systems have been analysed, there are even more queueing systems which have not been analysed yet. This does not entirely come as a surprise as existing results in queueing theory show that an apparently slight modification of the assumptions regarding arrivals, service times or service discipline, can often have a profound impact on the queueing dynamics, and therefore also on the performance of the queueing system. Therefore, despite the extensive body of literature on queueing systems, there is still a need for new efficient analysis methods.

This work focusses on a particular type of queueing models. We investigate queueing models that are comprised of multiple queues, where the customers in these different queues are served by a common service provider. Hence, each customer is not only affected by the presence of other customers in its own queue, but also by the presence of customers in all the other queues. To carefully capture
how customers in the different queues affect each other's performance, one cannot study a single queue in isolation. Instead, the queueing system should be studied as a single system with multiple queues which considerably complicates the analysis. We will adopt the term multidimensional queueing model for any such queueing system with multiple queues and shared service. For specific multidimensional queueing systems like particular multi-priority queueing systems, polling systems and processor sharing systems, analytical or numerical solutions are available. However, for the multidimensional queueing systems that are investigated in this work, such results were not available. For these systems, performance evaluation was practically conducted by means of simulation-based methods, while analytical or numerical solution would be more desirable to reduce the time it takes to conduct the performance analysis. This is most important if one wishes to dimension the queueing system. Optimal dimensioning being an optimisation problem and any optimisation routine requiring multiple evaluations of the performance measures for different parameter values, enabling fast calculation of the performance measures of interest also allows for fast performance optimisation.

The remainder of this introductory chapter is organised as follows. We first describe the details of the applications in the area of telecommunication networks and production systems that motivated the multidimensional queueing models that are analysed in this work. We then describe the theoretical background on Markov chains and their solution methods in section 1.3 . We discuss these solution methods in the context of multidimensional queueing systems in section 1.4 before outlining the remainder of the dissertation in section 1.5 . We list the publications that resulted from this work in section 1.6

### 1.2 Applications of multidimensional queueing models

The first step in assessing the performance of a real-life queueing system comprises the construction of a mathematical (queueing) model. The modelling process aims at identifying the key determinants that govern the queueing dynamics and affect the performance measures of interest. The modelling process in general leads to a set of simplifications to ensure reasonable model complexity, while avoiding oversimplification as this would lead to incorrect quantitative results. Ideally, we would like to achieve a perfect trade-off between model complexity and accuracy, aiming for a precise performance estimation via the simplest model possible. While from the vantage point of queueing analysis, one-dimensional Markovian queueing models are preferred as these are easier to analyse, many performance problems emerging in telecommunication networks $[2]-4]$, and manufacturing and assembly systems [5--7], can only be studied by multidimensional Markov models.

In the remainder of this section, we describe the details of the three main performance problems that motivated the multidimensional queueing systems studied in this dissertation: performance evaluation of opportunistic scheduling in wireless networks, of Drive-thru Internet in vehicular communication, and assembly from multi-product inventory systems.

### 1.2.1 Opportunistic scheduling in wireless networks

Recent advances in wireless communication systems include a number of rapidly developing technologies within the framework of current 4G and future 5G networks like cognitive radio, multiple input multiple output (MIMO) systems, and opportunistic scheduling. While absorbing state-of-the-art ideas and technologies along with introducing new services, the evolution of wireless networks is driven by the same main objectives from one generation to another: boosting the data rates, expanding the bandwidth, reducing latency, and minimising energy consumption. These central challenges have continuously motivated industrial and academic research in the area of wireless communications [8, 9].

One of the technologies, leading to overall performance improvement in wireless communication is opportunistic scheduling. This concept refers to a crosslayer media access control which utilises advanced features of the physical layer by accounting for the channel conditions when selecting what to send next. The basic scenario which utilises opportunistic scheduling considers communication between a base station or access point (AP) and multiple mobile devices or users as depicted in Figure 1.1. The AP sends to the users over a noisy channel, channel conditions differing for the different users. An ideal opportunistic scheduler prefers to send to the users with the best channel conditions, while it also accounts for fairness between its users. Utilising channel information, a scheduler can greatly improve its average throughput in a multi-user system compared to a channel-unaware scheduler like a round-robin scheduler which schedules in circular order and ignores channel variations.

While implemented only recently, the basic idea of channel-aware scheduling emerged in literature more than 20 years ago [10]. The first proposed opportunistic schedulers were MaxRate [10], MaxWeight [11] and Exp-rule [12]. These schedulers are easy to implement, yet already exhibit a notable performance increase in terms of overall throughput. Since then, a tremendous amount of new scheduling schemes have been designed to cope with the changing performance requirements for one wireless system over another. For a thorough description of the history and tendencies in channel-aware scheduling, we refer to [13] and the references therein. Currently, only few opportunistic schedulers are already implemented [14], however, future generations of wireless communication protocols are expected to employ this technology more extensively.


Figure 1.1: Illustration of wireless access point under varying transmission environment

One of the problems related to the design of opportunistic schedulers is to quantify their performance such that an appropriate scheduler can be chosen for a given application and its Quality of Service requirements. As any outstanding packets for AP to mobile user communication and for mobile user to AP communication are stored in buffers awaiting transmission, the performance assessment of opportunistic scheduling is basically a queueing performance problem. While the dynamics of the long-term buffer behaviour under varying channel assumptions can be studied by means of queueing theory, such an analysis is not straightforward. Indeed, opportunistic scheduling aims at finding the best trade-off between sending packets that have been waiting for a long time and packets that can be sent over the best channel. At least at the conceptual level, packets for the different mobile users are kept in separate buffers. Whenever a packet is selected for transmission, the choice not only affects the buffer of the packet that is selected, but all other buffers as well, as no packet from these buffers was selected. Even if the packets at the AP are stored in a single buffer, the scheduling accounts for the channel conditions of their destination. In either case, the performance analysis requires a multidimensional queueing model.

In Chapter 4 , we present such a multidimensional queueing model and its performance analysis. We make some simplifying assumptions to enable describing the queueing dynamics by means of a continuous-time Markov chain with finite state space. However, the modelling assumptions are sufficiently versatile to allow for (i) evaluating various buffer-size and channel-quality aware opportunistic schedulers, (ii) to assess the impact of time correlation of the channel conditions and (iii) to assess the impact of cross correlation of the channel conditions. The latter can seriously affect the performance of opportunistic schedulers as an op-
portunistic scheduler cannot optimise sending to the best channel if the different channels are good or bad at the same time.

While the proposed queueing model is a continuous-time Markov process with a finite state space, the state space is prohibitively large to allow for direct calculations of performance measures of interest. Moreover, the generator matrix lacks structural properties like being skip-free in one direction or reversibility to simplify the calculations (see section 1.3 below). Hence, in literature one often relies on simulations-based methods to obtain performance measures of interest. In this dissertation, we show that it is possible to calculate the performance measures at acceptable computational cost, in two parts of the parameter space: under light traffic conditions, when the AP load is low, and in overload conditions when the load at the AP is very high.

### 1.2.2 Drive-thru Internet in vehicular communication

Again in the field of telecommunication systems, another application under study in this dissertation focuses on a performance problem in the area of vehicle communications. The emerging concept of Cooperative Intelligent Transportation Systems (C-ITS) suggests a widespread adoption of information and communication technologies in diverse vehicular applications aimed at increasing transport safety, efficiency and comfort. C-ITS vehicles exchange information with each other as well as with the roadside infrastructure in a heterogeneous wireless networking environment.

In particular, we consider a scenario of vehicle-to-infrastructure communication referred to in the literature as Drive-thru Internet [15]. Drive-thru Internet provides access to internet services by installing wireless access points along the road. This access is especially useful in remote regions where cellular networks and urban internet resources might be unavailable. We focus on the performance of downlink traffic from a single AP, that is, data traffic from this AP to the moving vehicles that pass by. The coverage region of the AP is divided into several zones, each zone associated with particular data rates. Both the range of service and the length of the zones with equal data rate are defined by the technical characteristics of the AP and additional environmental factors that influence channel capacity along the road. As an example, in Figure 1.2 the range is divided in three parts, namely, an entry zone, a production zone and an exit zone in accordance with [16]. While the AP provides internet connectivity within the coverage area, due to the zone-dependent channel quality, the maximal throughput for users in both entry and exit zones is significantly lower compared to the production zone.

The Drive-thru Internet scenario resembles the earlier problem of opportunistic scheduling to some extent. Indeed, the AP has to send packets to multiple cars in different regions with different channel qualities, temporarily buffering packets


Figure 1.2: Drive-thru Internet scenario
at the AP prior to transmission. Again the packets for the different users are stored in separate buffers, if not physically than at least conceptually. In contrast to the opportunistic scheduler of the preceding section, the channel quality of a mobile user now changes due to movement of the mobile user. From the vantage point of performance analysis, this leads again to a queueing model with multiple dimensions. The Drive-thru scenario however is more complex as now also the position of the cars needs to be tracked.

In Chapter [5, a Markovian model for Drive-thru Internet is presented, which tracks the positions of the cars and the number of packets for these cars at the AP. Such a Markov model allows for evaluating schedulers for Drive-thru Internet. The scheduler now decides on what to send next based on the positions of the cars and the number of packets at the AP for these cars.

### 1.2.3 Assembly systems

A manufacturing process often consists of multiple operational steps, converting parts, raw materials or structural items into semi-final products, and then finally into end products.

In this dissertation, the performance of an assembly operation of semi-finished products into an end product or into another semi-finished product under uncertainty in demand and production times is investigated. The assembly takes parts from multiple inventories that offer temporary storage to smooth out uncertainty in the various production processes and are constantly replenished by in-house production facilities. Figure 1.3 shows an abstract representation of such an assembly process. Multiple part inventories are replenished by production processes, the parts itself being used to assemble the end product. Performance measures of interest include the throughput of the assembly system as well as the distribution and


Figure 1.3: Illustration of a manufacturing system with assembly line.
moments of the number of semi-finished parts in the different inventories.
If multiple part inventories are involved, it may be hard to allocate sufficient space for all inventories near the assembly line. In this case, the semi-finished parts can be collected into a specially designed container or kit and then delivered to the assembly line. This strategy is referred to as kitting and not only mitigates storage requirements at the assembly line, but also allows for reducing seek times during assembly as all semi-finished products can be readily located in the kit [17- -20$]$. As from a modelling perspective kit construction and assembly are equivalent, the models for studying the assembly operation can also be used to study the kitting process.

To model the assembly (or kitting) operation, one has to explicitly account for the state of the multiple inventories as it is sufficient that one of the inventories is empty to block the assembly operation. This in turn affects the performance of the assembly system, performance measures of interest including the throughput of the assembly system as well as the size of the part inventories. The need for a multidimensional model contrasts with most stochastic inventory models in literature [21, 22]. In Chapter 2 and 3, Markovian models of the assembly operation are investigated. In both chapters, the model accounts for the multiple inventories, which are replenished in accordance with independent Poisson processes, while the assembly time is exponentially distributed. In Chapter 3, semi-finished products additionally have a holding date, meaning that the products cannot be stored indefinitely in the inventory, but will perish after some time. This perishability is captured by abandonment processes from the different inventories. Food products are a prime example of perishable semi-finished products. However, perishable semi-finished products are also found in biochemical production, and in battery and semiconductor manufacturing [23].

### 1.3 Methodology

We introduce the main concepts of the theory of Markov chains, and their numerical solution techniques.

### 1.3.1 Markov processes

Stochastic processes can be used to study the behaviour of dynamical systems influenced by random factors. In this dissertation, Markov processes play a predominant role, as such processes combine versatility and analytical and/or numerical tractability.

Let $\left\{X_{t}, t \in \mathbb{R}^{+}\right\}$be a stochastic process taking values in a denumerable set $\mathcal{X}$. That is, $\left\{X_{t}, t \in \mathbb{R}^{+}\right\}$is a collection of $\mathcal{X}$-valued random variables indexed by an index $t \in \mathbb{R}^{+}$, which is usually interpreted as time. Such a process is a Markov process provided it possesses the 'memorylessness' or the Markov property [24],

$$
\begin{align*}
\operatorname{Pr}\left[X_{t}=x_{t} \mid X_{t_{0}}=x_{t_{0}}, X_{t_{1}}=x_{t_{1}}, \ldots, X_{t_{n}}\right. & \left.=x_{t_{n}}\right] \\
& =\operatorname{Pr}\left[X_{t}=x_{t} \mid X_{t_{n}}=x_{t_{n}}\right] \tag{1.1}
\end{align*}
$$

for $0 \leq t_{0}<t_{1}<t_{2}<\cdots<t_{n}<t$ and for $x_{s} \in \mathcal{X}$ for $s \in\left\{t_{0}, \ldots, t_{n}, t\right\}$. For a Markov process, the set $\mathcal{X}$ is referred to as the state space of the Markov process, an element of the state space being a state of the Markov process. The expression above states that the probability distribution of the future state of the process $X_{t}$ can be predicted given the current state $X_{t_{n}}$ and does not depend on all previous states. In other words, the past and the future are independent given the present.

From the Markov property we have,

$$
\begin{aligned}
& \operatorname{Pr}\left[X_{t_{1}}=x_{t_{1}}, \ldots, X_{t_{n}}=x_{t_{n}}\right]= \\
& \operatorname{Pr}\left[X_{t_{1}}=x_{t_{1}}\right] \prod_{m=2}^{n} \operatorname{Pr}\left[X_{t_{m}}=x_{t_{m}} \mid X_{t_{m-1}}=x_{t_{m-1}}\right]
\end{aligned}
$$

such that the joint distribution of the process at different time instants can be expressed in terms of the transition probabilities $\operatorname{Pr}\left[X_{t_{m}}=x_{t_{m}} \mid X_{t_{m-1}}=x_{t_{m-1}}\right]$. This implies that the process is completely characterised by the distribution of $X_{0}$ and the transition probabilities.

It is often more convenient to describe the Markov process in terms of transition rates, rather than in terms of transition probabilities. The transition rate $q_{i j}(t)$ from state $i$ to state $j$ at time $t \in \mathbb{R}^{+}$is defined as

$$
\begin{equation*}
q_{i j}(t)=\lim _{\Delta \rightarrow 0} \frac{\operatorname{Pr}\left[X_{t+\Delta t}=j \mid X_{t}=i\right]}{\Delta t}, i \neq j, i, j \in \mathcal{X} . \tag{1.2}
\end{equation*}
$$

In the remainder, we always assume that the Markov process is time homogeneous, meaning that the transition rates do not depend on time, $q_{i j}(t) \doteq q_{i j}$ for all $t \in \mathbb{R}^{+}$.

Let $\pi_{i}(t)=\operatorname{Pr}\left[X_{t}=i\right]$ denote the probability that the Markov process is in state $i$ at time $t$. The probabilities $\pi_{i}(t)$ satisfy the set of differential equations,

$$
\begin{equation*}
\frac{d \pi_{i}(t)}{d t}=-\pi_{i}(t) \sum_{j \in \mathcal{X}, j \neq i} q_{i j}+\sum_{j \in \mathcal{X}, j \neq i} \pi_{j}(t) q_{j i}, i \in \mathcal{X} \tag{1.3}
\end{equation*}
$$

It is often convenient to introduce the following matrix representation of this set of equations. Let $\boldsymbol{\pi}(t)$ denote the row vector with elements $\pi_{i}(t), i \in \mathcal{X}$, and let $\mathbf{Q}$ be the matrix,

$$
\mathbf{Q}=\left[q_{i j}\right]_{i, j \in \mathcal{X}},
$$

where the diagonal elements are equal to,

$$
\begin{equation*}
q_{i i}=-\sum_{j \neq i} q_{i j} . \tag{1.4}
\end{equation*}
$$

$\mathbf{Q}$ is the (possibly infinite) generator matrix of the Markov process. In view of these definitions, the set of differential equations can be written as,

$$
\frac{d \boldsymbol{\pi}(t)}{d t}=\boldsymbol{\pi}(t) \mathbf{Q}
$$

The main quantitative result that we aim to obtain from the Markov model is its stationary distribution,

$$
\boldsymbol{\pi} \doteq \lim _{t \rightarrow \infty} \boldsymbol{\pi}(t)
$$

The stationary distribution allows for studying the long-run dynamics of the model and can be found by solving the system of balance equations

$$
\begin{equation*}
\pi \mathrm{Q}=\mathbf{0} \tag{1.5}
\end{equation*}
$$

combined with normalisation condition,

$$
\begin{equation*}
\boldsymbol{\pi} \mathbf{1}^{\top}=1 \tag{1.6}
\end{equation*}
$$

Here $\mathbf{0}$ is a row vector of zeros, $\mathbf{1}$ is a row vector of ones of appropriate dimension and $\mathbf{1}^{\top}$ denotes the transpose of $\mathbf{1}$.

In this dissertation, the Markov processes represent queueing systems, the transitions corresponding to arrivals and departures of customers, and the state of the Markov process representing the number of customers in some subsystem of the system. Usually, the ultimate goal of solving a queueing system is to quantify the system behaviour under various loads. Once the stationary distribution $\pi$ is calculated, various performance measures of the system can be found directly. Let $X$ be distributed as the stationary distribution of the Markov process. Most performance measures can then be expressed in terms of the expectation of some function of $X$,

$$
\mathrm{E}[f(X)]=\sum_{i \in \mathcal{X}} \pi_{i} f(i)
$$

For example, if a function $g$ maps the state $X$ of the system on the total number of customers $Q$ in the system, $Q=g(X)$, the mean and variance of the number of customers in the system can be calculated as,

$$
\mathrm{E}[Q]=\sum_{i \in \mathcal{X}} \pi_{i} g(i), \quad \operatorname{var}[Q]=\sum_{i \in \mathcal{X}} \pi_{i} g(i)^{2}-\left(\sum_{i \in \mathcal{X}} \pi_{i} g(i)\right)^{2}
$$

### 1.3.2 Numerical methods for solving queueing models

Generally speaking, directly solving a system of $N$ equations (like equation (1.5) where the number of states in $\mathcal{X}$ equals the number of equations $N$ ) comes with a computational complexity not smaller than $N^{2}$. It leads to a practically unattainable computational demand for systems with a large state space. However, it is sometimes possible to benefit from particular structural properties of the system of equations (or of the corresponding generator matrix), thereby significantly reducing the numerical complexity. In this subsection we describe a number of common approaches and the corresponding queueing models that allow for such a reduction in numerical complexity.

### 1.3.2.1 Product-form solutions

For many queueing systems, the state space of the underlying Markov process is either $\mathcal{X}=\mathbb{N}^{K}$ for some fixed $K \in \mathbb{N}_{0}$, or a subset of $\mathbb{N}^{K}$. The state space is multidimensional and most often each dimension tracks the state of a particular subcomponent of the queueing system. For example, for a Jackson network (a network of M/M/1 queues with random routing, see [25]), each dimension tracks the number of customers in a particular queue.

The state of the Markov process is then a $K$-dimensional vector with nonnegative integer elements. For $\mathbf{x}=\left[x_{1}, \ldots, x_{K}\right] \in \mathcal{X}$, let $\pi_{\mathbf{x}}$ denote the stationary probability to be in state $\mathbf{x}$,

$$
\pi_{\mathbf{x}}=\operatorname{Pr}[\mathbf{X}=\mathbf{x}]
$$

Here $\mathbf{X}=\left[X_{1}, \ldots, X_{K}\right]$ is a random variable, distributed like the stationary distribution of the Markov process.

Now assume that the stationary distribution can be decomposed as follows,

$$
\pi_{\mathbf{x}}=B \prod_{k=1}^{K} \pi_{x_{k}}^{(k)}
$$

where $\left\{\pi_{y}^{(k)}, y \in \mathbb{N}\right\}$ is a (one-dimensional) probability distribution for each $k=$ $1, \ldots, K$ and where $B$ is a constant which ensures the normalisation of the joint distribution. Such a solution is referred to as a product-form solution as the joint
distribution of the system state can be written as the product of one-dimensional distributions.

Clearly, if there is a product-form solution, the number of unknown probabilities that need to be calculated is far smaller. Indeed, if there are $M$ states for each dimension, the total number of states is $M^{K}$, meaning that $M^{K}$ state probabilities need to be calculated. On the other hand, there are only $M K$ unknown probabilities in the product-form formulation and one unknown constant. While the system of equations for these unknown probabilities is no longer linear (as can be seen from plugging the product-form representation in the balance equations), it turns out that the numerical complexity for solving the balance equations can be greatly reduced. As it is most convenient to describe the numerical solution methods for multidimensional Markov processes with a product-form solution in terms of cooperating subsystems, we will do so below.

RCAT formulation The Markov process $\mathbf{X}$ models a finite set of subcomponents $S_{1}, \ldots, S_{K}$, the state of the $k$ th component being described by the $k$ th element of the vector $\mathbf{X}$. We now define the possible transitions of the different subcomponents. Consider the $k$ th subcomponent. All transitions are characterised by a label $a_{k}$ and by the departing state $x_{k}$ and arrival state $x_{k}^{\prime}$ of the $k$ th subcomponent corresponding to this transition. Active transitions additionally have a rate, denoted by $q\left(x_{k} \xrightarrow{a_{k}} x_{k}^{\prime}\right)$, whereas passive transitions do not have such a rate. Imposing that all transitions with the same label and of the same component are either active or passive, let $\mathcal{A}_{k}$ and $\mathcal{P}_{k}$ denote the sets of active and passive labels for the $k$ th subcomponent.

The Markov process is now defined such that whenever the $k$ th subcomponent is in state $x_{k}$, and there is an active transition to some other state $x_{k}^{\prime}$ with label $a_{k}$ and with rate $q\left(x_{k} \xrightarrow{a_{k}} x_{k}^{\prime}\right)$, then there is a transition with this rate for the complete Markov chain, where the transition not only invokes a change of the state of the $k$ th subcomponent, but also state changes in all other components where there are passive transitions with the same label.

The subcomponents cannot be directly studied in isolation, as a transition in one subcomponent triggers state changes of other subcomponents. The Reversed Compound Agent Theorem (RCAT) algorithm [26], roughly stated, calculates the rates of the passive transitions of all components, such that the components can be studied in isolation. The following theorem gives conditions such that the complete Markov chain has product form, see [26, 27].

Theorem 1. Given a set of cooperating subcomponents $S_{1}, \ldots, S_{K}$, assume that the following conditions are satisfied:

1. for all $k$, if $a \in \mathcal{A}_{k}$, then for each state $x_{k}$ of $S_{k}$, there is exactly one state $x_{k}^{\prime}$ such that $x_{k}^{\prime} \xrightarrow{a} x_{k}$;
2. for all $k$, if $a \in \mathcal{P}_{k}$, then for each state $x_{k}$ of $S_{k}$ there is exactly one state $x_{k}^{\prime}$ such that $x_{k} \xrightarrow{a} x_{k}^{\prime}$;
3. there exists a set of positive values $\mathcal{K}=\left\{K_{a}, a \in \mathcal{A}_{k} \cup \mathcal{P}_{\ell}, k, \ell=1, \ldots, K\right\}$ such that when all subcomponents are closed using these values (closing means that the rates of the passive transitions are fixed), we have that $K_{a}$ is the rate of all transitions labeled by $a$ in the reversed processes of all subcomponents.

Then, the stationary distribution of each positive recurrent state $\mathbf{x}$ is in product form,

$$
\pi_{\mathbf{x}}=B \prod_{k=1}^{K} \pi_{x_{k}}^{(k)}
$$

where $\pi^{(k)}$ is the stationary distribution of the $k$ th component (after closing).
The above method establishes existence of a product-form solution. The Iterative Numerical Algorithm for Product Forms (INAP) [26, 27] allows for calculating the solution. The base version of INAP operates as follows:

1. Initialisation: $f \leftarrow 0$, set up randomly $\pi^{(k)}[0]$ for all $k=1, \ldots, K$;
2. For all synchronising transitions $a$ (a transition is synchronising if it invokes a transition in another subcomponent), compute $K_{a}[f]$ as the mean of the reversed rates of the transitions labelled by $a$, using $\pi^{(k)}[f]$ with $k$ such that $a \in \mathcal{A}_{k}$;
3. For all $k=1, \ldots, K$, close the subcomponents with the rates $K_{a}$ found in 3, for all all $a \in \mathcal{P}_{k}$;
4. $f \leftarrow f+1$;
5. For all $k=1, \ldots, K$, compute $\pi^{(k)}[f]$ as the stationary solution of $S_{k}$ (after closing);
6. If there exists a $k \in\{1, \ldots, K\}$ such that $\pi^{(k)}[f] \neq \pi^{(k)}[f-1]$ within precision $\epsilon$ and $f \leq T$, cycle to step 2 .
7. Terminate with one of the following options:

- If $f>T$ return "No product-form solution found";
- For all synchronising transitions $a$, use $\pi^{(k)}[f]$ with $k$ such that $a \in$ $\mathcal{A}_{k}$ to check if the reversed rates of all transitions labelled with $a$ are constant. If this is the case for all $k$, return the product-form solution $\prod_{k} \pi^{(k)}[f]$. If this is not the case, return "No product-form solution found".

In the above $T$ is the maximal number of iterations and $\epsilon$ is the precision.

Jackson Networks To illustrate how the availability of a product-form solution can simplify the solution, we consider a Jackson network, a queueing network proposed by J.R. Jackson [25] for analysing the operation of a job shop. A more recent application of such networks can be found in [20], in which the authors investigate the performance of a flexible manufacturing system that can process various jobs synchronously. For Jackson networks, a solution is available in closed form, hence there is no need to apply the INAP algorithm.

A Jackson network is a network of $K>0$ queues, such that (i) any external arrivals to the different queues constitute Poisson processes, (ii) all service times are exponentially distributed and (iii) the service discipline at all queues is firstcome, first-served. Finally, (iv) upon service completion at queue $k$, the customer joins queue $\ell$ with probability $p_{k \ell}$ or leaves the system with probability

$$
\bar{p}_{k}=1-\sum_{\ell=1}^{K} p_{k \ell} .
$$

This is a so-called open network, there are also variants which are closed or semiopen, see [28]. In a closed network there are neither external arrivals nor departures from the system, such that a fixed number of customers move from queue to queue. Semi-open networks combine features of both open and closed networks by allowing external arrivals and departures but also by limiting the total number of customers in the system.

To find the solution of the network, one first calculates the arrival rates $\theta_{k}$ at the different queues $(k=1, \ldots, K)$. The arrival rate not only includes the external arrival rate to the queue $\lambda_{k}$, but also the rate at which customers from other queues arrive at the $k$ th queue. Noting that all customers that enter the queue, also eventually leave the queue, one finds the following set of equations for the arrival rates,

$$
\theta_{k}=\lambda_{k}+\sum_{\ell=1}^{K} \theta_{\ell} p_{\ell k}
$$

Let $\mu_{k}(n)$ denote the service rate at the $k$ th queue, when there are $n$ customers in this queue. Given the different $\theta_{k}$, the product-form solution then reads,

$$
\pi(\mathbf{x})=\prod_{k=1}^{K} \pi_{x_{k}}^{(k)}
$$

with,

$$
\pi_{x}^{(k)}=\frac{1}{\sum_{y=0}^{\infty} \frac{\left(\theta_{k}\right)^{y}}{\prod_{n=1}^{y} \mu_{k}(n)}} \frac{\left(\theta_{k}\right)^{x}}{\prod_{n=1}^{x} \mu_{k}(n)},
$$

provided that the sum in the denominator of the first factor of the right hand side converges. If this sum does not converge for some $k$, the queueing network is not stable.

### 1.3.2.2 Matrix-geometric method

Consider a Markov process $\mathbf{X}(t)=[L(t), P(t)]$ with a two-dimensional state space. Here, $L(t)$ is the level and $P(t)$ is the phase of the process $\mathbf{X}(t)$. Now assume that the system only takes a finite number of possible phases, and that transitions between levels are skip-free. This means that one can only have transitions to adjacent levels. By ordering the states according to levels first, one sees that the generator matrix of the Markov process has the following block representation,

$$
\mathbf{Q}=\left(\begin{array}{cccccc}
\mathbf{B}_{1} & \mathbf{B}_{2} & & & & \\
\mathbf{A}_{10} & \mathbf{A}_{11} & \mathbf{A}_{12} & & & \\
& \mathbf{A}_{20} & \mathbf{A}_{21} & \mathbf{A}_{22} & & \\
& & \mathbf{A}_{30} & \mathbf{A}_{31} & \mathbf{A}_{32} & \\
& & & \ddots & \ddots & \ddots
\end{array}\right)
$$

where $\mathbf{B}_{2}, \mathbf{A}_{i 0}$, and $\mathbf{A}_{i 2}$ are matrices with non-negative entries and where the non-diagonal entries of $\mathbf{B}_{1}$ and $\mathbf{A}_{i 1}$ are non-negative as well. Finally, the diagonal entries of $\mathbf{B}_{1}$ and $\mathbf{A}_{i 1}$ are negative such that the row sums of $\mathbf{Q}$ are zero. Such Markov processes are referred to as Quasi-Birth-Death processes (QBD), as there are only transitions between adjacent levels, like there are only transitions between adjacent states in birth-death processes.

We focus in particular on QBDs which are homogeneous in the following way,

$$
\mathbf{Q}=\left(\begin{array}{cccccc}
\mathbf{B} & \mathbf{A}_{2} & & & & \\
\mathbf{A}_{0} & \mathbf{A}_{1} & \mathbf{A}_{2} & & & \\
& \mathbf{A}_{0} & \mathbf{A}_{1} & \mathbf{A}_{2} & & \\
& & \mathbf{A}_{0} & \mathbf{A}_{1} & \mathbf{A}_{2} & \\
& & & \ddots & \ddots & \ddots
\end{array}\right)
$$

For such systems, the stationary solution can be efficiently calculated by means of the matrix geometric method [29]. Some applications of Markovian queues which are represented by QBDs include queueing systems with Markovian arrival processes including the interrupted Poisson process, multiserver queues with phasetype distributed service times, queues with a traffic shaper, priority queues where one class has a finite buffer, etc.

For ease of notation, consider the following block representation of the stationary distribution

$$
\begin{equation*}
\boldsymbol{\pi}=\left[\boldsymbol{\pi}_{0}, \boldsymbol{\pi}_{1}, \boldsymbol{\pi}_{2}, \ldots\right] \tag{1.7}
\end{equation*}
$$

where

$$
\begin{equation*}
\boldsymbol{\pi}_{k}=[\pi(k, 1), \ldots, \pi(k, M)] \tag{1.8}
\end{equation*}
$$

with

$$
\pi(k, m)=\lim _{t \rightarrow \infty} \operatorname{Pr}[L(t)=k, P(t)=m]
$$

for $m=1, \ldots, M$. Here $M$ represents the number of possible phases (such that the matrices $\mathbf{A}_{0}, \mathbf{A}_{1}, \mathbf{A}_{2}$, and $\mathbf{B}$ are $M \times M$ matrices).

For the block representations of the stationary probability vector and the generator matrices, one directly obtains the following set of equations,

$$
\begin{array}{r}
\boldsymbol{\pi}_{0} \mathbf{B}+\boldsymbol{\pi}_{1} \mathbf{A}_{0}=0, \\
\boldsymbol{\pi}_{k-1} \mathbf{A}_{2}+\boldsymbol{\pi}_{k} \mathbf{A}_{1}+\boldsymbol{\pi}_{k+1} \mathbf{A}_{0}=0, \tag{1.10}
\end{array}
$$

for $k>0$. As transitions between the levels of the QBD are invariant to the level, one expects that the solution takes the form,

$$
\begin{equation*}
\boldsymbol{\pi}_{k}=\boldsymbol{\pi}_{k-1} \mathbf{R}=\boldsymbol{\pi}_{0} \mathbf{R}^{k} \tag{1.11}
\end{equation*}
$$

for $k>0$ and some unknown $M \times M$ rate matrix $\mathbf{R}$. If one now plugs the proposal solution 1.11 into 1.10 , one finds that the rate matrix $\mathbf{R}$ is the minimal non-negative solution of the quadratic matrix equation

$$
\begin{equation*}
\mathbf{A}_{2}+\mathbf{R} \mathbf{A}_{1}+\mathbf{R}^{2} \mathbf{A}_{0}=\mathbf{0} \tag{1.12}
\end{equation*}
$$

This equation suggests the following iterative algorithm to calculate $\mathbf{R}$. Start with $\mathbf{R}=\mathbf{0}$, the $M \times M$ matrix with all entries equal to zero. Then update the matrix as follows,

$$
\mathbf{R} \leftarrow-\left(\mathbf{A}_{2}+\mathbf{R}^{2} \mathbf{A}_{0}\right) \mathbf{A}_{1}^{-1},
$$

till there is convergence. This recursion indeed converges provided the QBD process is an ergodic Markov process. An irreducible QBD process is ergodic if the following stability condition holds [30],

$$
\begin{equation*}
\boldsymbol{\pi}_{A} \mathbf{A}_{0} \mathbf{1}^{\top}<\boldsymbol{\pi}_{A} \mathbf{A}_{2} \mathbf{1}^{\top} \tag{1.13}
\end{equation*}
$$

where row vector $\boldsymbol{\pi}_{A}$ satisfies

$$
\boldsymbol{\pi}_{A}\left(\mathbf{A}_{0}+\mathbf{A}_{1}+\mathbf{A}_{2}\right)=\mathbf{0}
$$

This stability condition means that the drift of the process to higher levels is smaller than the drift to the lower levels. In addition, if the QBD process is ergodic, it can be shown that the largest eigenvalue of $\mathbf{R}$ is less than one, such that

$$
\sum_{k=0}^{\infty} \mathbf{R}^{k}=(\mathbf{I}-\mathbf{R})^{-1}
$$

where $\mathbf{I}$ denotes the $M \times M$ identity matrix.
Combining (1.9) and 1.11 , the remaining unknown vector $\pi_{0}$ adheres,

$$
\begin{equation*}
\boldsymbol{\pi}_{0} \mathbf{B}+\boldsymbol{\pi}_{0} \mathbf{R} \mathbf{A}_{0}=\mathbf{0} \tag{1.14}
\end{equation*}
$$

This equation only allows for determining $\pi_{0}$ up to a normalising constant. The normalisation condition $\boldsymbol{\pi} \mathbf{1}^{\top}=1$ combined with (1.11) finally yields

$$
\begin{equation*}
\sum_{k=0}^{\infty} \boldsymbol{\pi}_{k} \mathbf{1}^{\top}=\sum_{k=0}^{\infty} \boldsymbol{\pi}_{0} \mathbf{R}^{k} \mathbf{1}^{\top}=\boldsymbol{\pi}_{0}(\mathbf{I}-\mathbf{R})^{-1} \mathbf{1}^{\top}=1 \tag{1.15}
\end{equation*}
$$

such that this normalising constant can be determined.

### 1.3.2.3 Iterative methods

Both Markov processes with product-form solutions and QBD processes provide computationally attractive solution techniques. However, both types of processes impose hefty structural properties on the Markov processes which are often not met for practical applications. In many realistic applications product form does not hold due to types of interdependencies between the different queues. Similarly, for many problems, the state space cannot be ordered such that the generator matrix takes a block diagonal form. In cases where the Markov processes do not allow for a simplified solution, the stationary distribution must be found by directly solving the system of balance equations (1.5). A direct solution of the system of linear equations is most often computationally prohibitive. For example, Gaussian elimination exhibits a numerical complexity of order $O\left(N^{3}\right)$, with $N$ the number of unknowns. To mitigate the computational demands of exact methods, one can rely on iterative methods instead. The order of complexity for stationary iterative methods is $O\left(N^{2}\right)$ for each iteration and additional gain in computation time can be achieved by the parallel implementation. Moreover, the generator matrix is often sparse, the number of non-zero entries in the generator matrix being far smaller than $N$. If the number of non-zero elements does not grow with $N$, the computational complexity for a single iteration is just $O(N)$, meaning that iterative methods can be very fast provided not too many iterations are required.

Iterative methods for solving large systems of linear equations mostly rely on projection techniques to obtain an approximate solution. Here we review basic iterative methods that are applied in this dissertation: Jacobi, Gauss-Seidel and Successive Over-Relaxation (SOR) [31]. We discuss methods for a system of equations

$$
\begin{equation*}
\mathbf{A x}=\mathbf{b} \tag{1.16}
\end{equation*}
$$

where $\mathbf{A}$ is a real matrix of size $N \times N$, and $\mathbf{b} \in \mathbb{R}^{N}$ and $\mathbf{x} \in \mathbb{R}^{N}$ are column vectors with $N$ entries. For further use, we also introduce some notation for the decomposition of $\mathbf{A}$ into a matrix with A's diagonal elements $\mathbf{D}$, and a matrix $\mathbf{R}$ with A's off-diagonal elements $\mathbf{R}$,

$$
\mathbf{A}=\mathbf{D}+\mathbf{R}
$$

In addition, $\mathbf{R}$ can be further decomposed into a matrix $\mathbf{E}$ with the lower triangular entries of $\mathbf{R}$ and a matrix $\mathbf{F}$ with the upper triangular entries of $\mathbf{R}$,

$$
\mathbf{R}=\mathbf{E}+\mathbf{F} .
$$

All the considered iterative methods calculate consecutive proposal solutions $\mathbf{x}^{(r)}, r=0,1, \ldots$ of the system of equations (1.16). One starts with an initial solution $\mathbf{x}^{(0)}$. Given the $r$ th solution, the $(r+1)$ st solution $\mathbf{x}^{(r+1)}$ is calculated using the results from the preceding iteration. The exact way to recalculate the next proposal solution differentiates the different solution methods that are discussed below. All iterative methods aim at annihilating the elements of the residual vector $\mathbf{b}-\mathbf{A} \mathbf{x}^{(r)}$ for increasing $r$. Of course, once the residual vector converges to $\mathbf{0}$ for some $r, \mathbf{x}^{(r)}$ is the solution of the matrix equation 1.16. Typically, the iterative procedure however only converges for an infinite number of iterations. Hence, the procedure is terminated once the elements of the residual vector are sufficiently small. An example of a convergence criteria is given below

$$
\begin{equation*}
\frac{\left\|\mathbf{x}^{(r)}-\mathbf{x}^{(r-1)}\right\|_{1}}{\left\|\mathbf{x}^{(r)}\right\|_{1}}<\epsilon \tag{1.17}
\end{equation*}
$$

where threshold $\epsilon$ is chosen to ensure the precision of the approximate solution and $\|\cdot\|_{1}$ denotes the $\ell_{1}$ norm.

Before introducing the different procedures, note that it is easy to write the balance equations, combined with the normalisation condition in the form 1.16). Let 1 be a row vector of ones, we then have from (1.5) and (1.6),

$$
\left(\mathbf{Q}+\mathbf{1}^{\top} \mathbf{1}\right)^{\top} \boldsymbol{\pi}^{\top}=\mathbf{1}^{\top}
$$

where $\mathbf{x}^{\top}$ denotes the transpose of $\mathbf{x}$.

Jacobi method The simplest iterative procedure is the Jacobi method. This method is easy to implement and exhibits low complexity. Jacobi iterations aim to reduce the residual vector as follows,

$$
\begin{equation*}
\mathbf{x}^{(r+1)}=\mathbf{D}^{-1}\left(\mathbf{b}-\mathbf{R} \mathbf{x}^{(r)}\right) . \tag{1.18}
\end{equation*}
$$

In component-wise form, this can be written as,

$$
\begin{equation*}
x_{n}^{(r+1)}=\frac{1}{a_{n n}}\left(b_{n}-\sum_{m \neq n} a_{n m} x_{m}^{(r)}\right) \tag{1.19}
\end{equation*}
$$

for $n=1, \ldots, N$ with $a_{n m}$ the $(n, m)$ th element of $\mathbf{A}$ and $x_{n}^{(r)}$ the $n$th element of $\mathbf{x}^{(r)}$. The Jacobi method converges to the exact solution, provided the spectral
radius of $\mathbf{D}^{-1} \mathbf{R}$ is less than one. A sufficient condition for convergence is that $\mathbf{A}$ is strictly diagonally dominant. That is, the absolute values of the elements $a_{n m}$ satisfy the inequalities,

$$
\left|a_{n n}\right|>\sum_{m \neq n}\left|a_{n m}\right|,
$$

for $n=1, \ldots, N$.

Gauss-Seidel method The Gauss-Seidel method is similar to the Jacobi method, the only difference being that the Gauss-Seidel method uses the already updated elements of $\mathbf{x}^{(r+1)}$, while calculating further elements of $\mathbf{x}^{(r+1)}$. In matrix form, the Gauss-Seidel method calculates $\mathbf{x}^{(r+1)}$ as follows,

$$
\begin{equation*}
\mathbf{x}^{(r+1)}=\mathbf{D}^{-1}\left(\mathbf{b}-\mathbf{E} \mathbf{x}^{(r+1)}-\mathbf{F} \mathbf{x}^{(r)}\right) \tag{1.20}
\end{equation*}
$$

In element-wise form, we have,

$$
\begin{equation*}
x_{i}^{(r+1)}=\frac{1}{a_{i i}}\left(b_{i}-\sum_{j=1}^{i-1} a_{i j} x_{j}^{(r+1)}-\sum_{j=i+1}^{n} a_{i j} x_{j}^{(r)}\right), \tag{1.21}
\end{equation*}
$$

for $i=1, \ldots, N$. Compared to the Jacobi method, the elements of $\mathbf{x}$ can be overwritten as they are computed in this algorithm. Therefore, only one storage vector is needed. The Gauss-Seidel method converges when $\mathbf{A}$ is symmetric positivedefinite or strictly diagonally dominant. Both are sufficient but not necessary conditions for convergence.

Successive Over-Relaxation method Finally, in the SOR algorithm, the iteration takes the form

$$
\begin{equation*}
\mathbf{x}^{(r+1)}=\omega \mathbf{D}^{-1}\left(\mathbf{b}-\mathbf{E} \mathbf{x}^{(r+1)}-\mathbf{F} \mathbf{x}^{(r)}\right)+(1-\omega) \mathbf{x}^{(r)} \tag{1.22}
\end{equation*}
$$

where a pre-selected relaxation parameter $\omega$ is chosen to accelerate or stabilise convergence. When $\omega=1$, the SOR method reduces to the Gauss-Seidel method. The case $\omega<1$ corresponds to under-relaxation and can be applied if the GaussSeidel method does not converge. The case $\omega>1$ refers to over-relaxation and allows for accelerating convergence in comparison with the Gauss-Seidel method.

Relaxation can also be applied on the Jacobi method, which then leads to the weighted Jacobi method

$$
\begin{equation*}
\mathbf{x}^{(r+1)}=\omega \mathbf{D}^{-1}\left(\mathbf{b}-\mathbf{R} \mathbf{x}^{(r)}\right)+(1-\omega) \mathbf{x}^{(r)}, \tag{1.23}
\end{equation*}
$$

a typical choice for $\omega$ being $\omega=2 / 3$.

### 1.3.2.4 Series expansion techniques

Series expansion techniques allow for calculating the Taylor series expansion of the stationary distribution of the Markov process in some parameter of the system. Rather than calculating the solution for a single parameter setting, series expansion techniques can be used to calculate the stationary distribution where one parameter is allowed to vary in a region around some fixed parameter value. Series expansion techniques for Markov processes are sometimes referred to as perturbation techniques, the power series algorithm or light-traffic approximations. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of performance measures with respect to the system parameters. In particular singular perturbations where the perturbation does not preserve the class-structure of the non-perturbed chain, have received considerable attention in literature, see [32-34] and the references therein. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by an auxiliary variable $\varepsilon$. For $\varepsilon=0$, the chain can be solved efficiently, and one can also obtain the perturbation of the chain in $\varepsilon$. For $\varepsilon=1$, the original Markov chain is retrieved such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of the series expansion includes $\varepsilon=1$, see e.g. [35-38]. Finally, light-traffic approximations often correspond to a series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in stochastic systems, we further refer the reader to the surveys in [39] and [40]. We focus on regular perturbations below, and indicate how such perturbation techniques can be used to efficiently study Markov processes.

Regular perturbation Consider a Markov process with finite state space and generator matrix $\boldsymbol{Q}_{\varepsilon}$, where we added the subscript to make its dependence on the parameter $\varepsilon$ explicit. We assume that the transition rates are linear functions of the parameter $\varepsilon$, such that the generator matrix can be written as,

$$
\begin{equation*}
\mathbf{Q}_{\varepsilon}=\mathbf{Q}^{(0)}+\varepsilon \mathbf{Q}^{(1)}, \tag{1.24}
\end{equation*}
$$

where neither $\mathbf{Q}^{(0)}$ nor $\mathbf{Q}^{(1)}$ depends on the perturbation parameter $\varepsilon$. These matrices represent the unperturbed and perturbed parts, respectively. The steady-state probability vector function obviously also depends on $\varepsilon$; we write $\boldsymbol{\pi}(\varepsilon)$, such that the balance equations read,

$$
\begin{equation*}
\boldsymbol{\pi}(\varepsilon) \mathbf{Q}_{\varepsilon}=\boldsymbol{\pi}(\varepsilon)\left(\mathbf{Q}^{(0)}+\varepsilon \mathbf{Q}^{(1)}\right)=\mathbf{0} \tag{1.25}
\end{equation*}
$$

We now consider the series expansion around some value $\varepsilon_{0}$. To this end, we introduce the Taylor series expansion,

$$
\begin{equation*}
\boldsymbol{\pi}(\varepsilon)=\sum_{n=0}^{\infty} \boldsymbol{\pi}_{n}\left(\varepsilon-\varepsilon_{0}\right)^{n} \tag{1.26}
\end{equation*}
$$

Note that by Cramer's rule, one directly finds that the elements of $\boldsymbol{\pi}(\varepsilon)$ are rational functions of $\varepsilon$. If the Markov process is ergodic for $\varepsilon=\varepsilon_{0}$, the elements of $\boldsymbol{\pi}(\varepsilon)$ are finite. Hence, the elements of $\boldsymbol{\pi}(\varepsilon)$ have no pole in $\varepsilon=\varepsilon_{0}$, which implies that the elements of $\boldsymbol{\pi}(\varepsilon)$ are analytic functions in a neighbourhood of $\epsilon_{0}$. This in turn justifies the series expansion.

We focus on regular perturbations. In this case, $\mathbf{Q}_{\varepsilon_{0}}$ is the generator matrix of a Markov process with a single ergodic class. Plugging the series expansion 1.26) into (1.25), we get

$$
\begin{equation*}
\boldsymbol{\pi}(\varepsilon) \mathbf{Q}_{\varepsilon}=\sum_{n=0}^{\infty} \boldsymbol{\pi}_{n}\left(\varepsilon-\varepsilon_{0}\right)^{n}\left(\mathbf{Q}^{(0)}+\varepsilon_{0} \mathbf{Q}^{(1)}+\left(\varepsilon-\varepsilon_{0}\right) \mathbf{Q}^{(1)}\right)=\mathbf{0} \tag{1.27}
\end{equation*}
$$

We can now identify the terms with equal powers of $\left(\varepsilon-\varepsilon_{0}\right)$. This leads to the following set of equations for the terms in the series expansion of $\boldsymbol{\pi}(\epsilon)$,

$$
\boldsymbol{\pi}_{0}\left(\mathbf{Q}^{(0)}+\varepsilon_{0} \mathbf{Q}^{(1)}\right)=\mathbf{0}, \quad \boldsymbol{\pi}_{n+1}\left(\mathbf{Q}^{(0)}+\varepsilon_{0} \mathbf{Q}^{(1)}\right)=-\boldsymbol{\pi}_{n} \mathbf{Q}^{(1)}
$$

for $n=0,1, \ldots$ Complementing the former set of equations with the normalisation conditions,

$$
\begin{equation*}
\boldsymbol{\pi}_{0} \mathbf{1}^{\top}=1, \quad \boldsymbol{\pi}_{n} \mathbf{1}^{\top}=0 \tag{1.28}
\end{equation*}
$$

allows for recursive calculation of the terms $\boldsymbol{\pi}_{n}$, for $n>0$.
From a computational point of view, the equations for calculating the terms in the series expansion resemble the balance equations. Hence, in the general case, the numerical complexity for calculating the terms in the series expansion is of the same order as the calculation of the stationary vector. As such, the method does not simplify the complexity of the calculation, but rather shows that the series expansion can be calculated with the same complexity. In this dissertation, we rely on the iterative procedures of the preceding section to solve the equations for the different terms in the series expansion.

For many systems however, it turns out that the series expansion is considerably easier to calculate, at least for particular $\varepsilon_{0}$-values. This is the case, if the matrix $\left(\mathbf{Q}^{(0)}+\varepsilon_{0} \mathbf{Q}^{(1)}\right)$ is upper or lower triangular. In these cases, the matrix equation can be solved in just $O(N)$ (assuming the matrix is sparse as well), with $N$ the size of the state space. In other words, assuming that the generator matrix is sparse, we get the series expansion at the computational cost of getting the value in a single point by an iterative procedure.

Multiple expansions For a fixed $\varepsilon_{0}$, the series expansion only converges to the correct value in a limited region around $\varepsilon_{0}$. The region of convergence is an open disk, centred in $\varepsilon_{0}$, and with radius equal to the distance between $\varepsilon_{0}$ and the closest (complex) pole of $\boldsymbol{\pi}(\varepsilon)$. Note that there is no pole for all $\varepsilon$ for which the Markov process is ergodic, as then $\boldsymbol{\pi}(\varepsilon)$ is well defined. If the Markov process is ergodic


Figure 1.4: Illustration of the convergence radius of two series expansions of a function with a complex pole. Convergence regions $R_{0}$ and $R_{1}$ of Taylor series for $\varepsilon=0$ and $\varepsilon=\varepsilon_{1}$ correspondingly of a function with a complex pole.
for all non-negative $\varepsilon$, this means that the poles are either complex or real and negative. No matter how many terms we calculate in the series expansion, there is no convergence to the correct value beyond this region of convergence.

Therefore it makes sense to consider the series expansion in multiple points, say $\varepsilon_{0}, \varepsilon_{1}$, etc. One would prefer to choose the next value within the region of convergence of one of the preceding values. In this case, the series expansion in the preceding value can be chosen as an estimate of the next value. As this estimate is already accurate, only a few iterations will be needed to calculate the terms in the next point. However, in many practical applications, the position of the poles of $\boldsymbol{\pi}(\varepsilon)$ prevents one to considerably extend the region of convergence by calculating the terms in a point within the region of convergence. This is most clearly the case if the closest pole is close to the real axis, at the side where one aims at extending the region of convergence.

The problem of choosing the value for the next series is illustrated in Figure 1.4 The first series expansion is around $\varepsilon=0$ (this is a Maclaurin series expansion), and has convergence radius $R_{0}$, which is equal to the modulus of the complex pole with the smallest absolute value. The position of this pole is indicated on the figure as well. To extend the region of convergence, a new series expansion is calculated around $\varepsilon=\varepsilon_{1}$. The convergence radius $R_{1}$ is determined by the pole which is closest to $\varepsilon_{1}$. In the figure, this is the same pole that determined the radius of convergence of the first expansion. Of course, it is possible that other poles are closer to $\varepsilon_{1}$, in which case the regions of convergence of the expansions may not overlap. If one would choose $\varepsilon_{1}$ within the region of convergence of the expansion around $\varepsilon=\varepsilon_{0}$, the extension of the region of convergence is at most the indicated distance $\Delta$ between $R_{0}$ and the complex pole.

Performance measures. Once the terms of the expansion $\boldsymbol{\pi}_{0}, \boldsymbol{\pi}_{1}, \ldots, \boldsymbol{\pi}_{N}$ are found, we can again calculate various performance measures. Let $X_{\varepsilon}$ be a random variable with stationary distribution $\pi(\varepsilon)$, let $\mathcal{X}$ be the corresponding state space, and let $f: \mathcal{X} \rightarrow \mathbb{R}$ be a deterministic function, then we have

$$
\mathrm{E}\left[f\left(X_{\varepsilon}\right)\right]=\sum_{n=0}^{N} \sum_{i \in \mathcal{X}} \pi_{n}(i)\left(\varepsilon-\varepsilon_{0}\right)^{n} f(i),
$$

where (with a slight abuse of notation) $\pi_{n}(i)$ denotes the $n$th component in the series expansion around $\varepsilon=\varepsilon_{0}$ of the stationary probability to be in state $i$.

### 1.4 Multidimensional queueing models

In this subsection we discuss the model assumptions and numerical challenges related to the multidimensional queueing systems studied in this work. Generally speaking multidimensionality of the stochastic process appears if the model under study combines several individual stochastic processes. In this work, the subcomponents are mostly Markovian queues (queues with Poisson arrivals and exponential service times), although sometimes an additional dimension is introduced for a modulating background process. Such processes are Markov processes with a limited number of states which modulate the parameters of the other queueing processes. More precisely, in each state of the modulating process, a fixed set of parameters of the other processes can be chosen, see also Chapters 3 and 4

The aggregated state space of the combined process is the product space of the state spaces of the different subcomponents. If there are $K$ such components, each component being in one of $N$ possible states, the size of the aggregated state space is $N^{K}$. As there might be a transition from any state to any state, there can be $N^{K}(N-1)^{K}$ transitions. Luckily, the number of possible transitions is far less in the models considered in this work. For each state, the number of possible transitions is typically of the order $O(K)$, which means that the transition matrix is sparse. For such sparse matrices, iterative solution techniques can be readily applied. We however will combine such iterative techniques with series expansion techniques and identify numerous cases in which the numerical complexity of calculating a single term in the series expansion has the computational complexity of a single iteration in iterative solution techniques. In addition, some of the numerical methods will benefit from identifying a block matrix representation of the generator matrix, like in the matrix geometric methods. To illustrate our contributions, we discuss the numerical approach for two of the queueing systems which will be investigated in detail in the following chapters.

Coupled queueing As an example of a multidimensional queueing model, consider a system with $K$ queues and coupled service as depicted in Figure 1.5. Ser-


Figure 1.5: Queueing model for assembly systems.


Figure 1.6: State space size of the coupled queueing system vs. the queue capacity $C$ for different numbers of queues (left) and vs. the number of queues $K$ for different queue capacities (right).
vice coupling means that customer departures from the queues are synchronised, service is provided for a batch of customers, one customer from each queue. Moreover, once one of the queues is empty, there is no service. This is a queueing model for assembly operations with in-house production and is discussed in detail in Chapter 2 and Chapter 3. The queues represent inventories of semi-finished products that supply the assembly line. Production is only possible if all inventories are nonempty, meaning that the inventories are coupled.

Assuming that all arrivals occur in accordance with a Poisson process and that service times are exponentially distributed, the state of the system is described by the number of customers in the different queues. If all queues have capacity $C$ and there are $K$ queues, the size of the state space is $(C+1)^{K}$. Clearly, the size of the state space grows very fast, both in $C$ and $K$. This is also illustrated in Figure 1.6, which depicts the state space size of the coupled queueing system vs. the queue capacity $C$ for different numbers of queues and vs. the number of queues $K$ for different queue capacities. In contrast, from any state there are at most $K+1$ possible transitions: $K$ transitions correspond to arrivals in the different queues, and 1 transition corresponds to a service completion. Of course, when some queues
are empty, there is no service, and when a queue is full, there cannot be arrivals. In other words, the number of transitions from a state is $O(K)$ and the generator matrix of the process is sparse.

As the generator matrix is sparse, we can readily apply iterative solution techniques. Moreover, if the service rate is sent to 0 , the resulting generator matrix has additional structure: assuming the states are ordered lexicographically, there are only upward transitions (such that the generator matrix is upper triangular). By this observation, it is possible to calculate the terms in the series expansion around service rate 0 in $O\left(C^{K}\right)$ (the order of the state space). Expansions around other service rates do not have this structural property. For these expansions, we use iterative methods to calculate the terms in the series expansion.

Opportunistic scheduling As another example, we briefly describe a model for downlink traffic in a multi-user wireless access point scenario with a channelaware scheduler and multiple mobile nodes. The model is extensively studied in Chapter 4. The queueing model consists of $K$ queues as depicted in Figure 1.7. each queue holding the packets destined for a particular mobile node. The packets in all queues arrive in accordance with Poisson processes and have exponentially distributed packet sizes. The transmitter of the wireless access point operates an opportunistic scheduling policy, meaning that the transmitter is aware of the channel conditions of the different mobile nodes and can choose which node it sends to next, depending on the channel and queueing conditions. Introducing a background process for the channel conditions - this is a finite Markov process such that there are different channel conditions in each state - the transmission rate for the $k$ th queue $\mu_{k}(m, \mathbf{i})$ is a function of the state $m$ of the background process of the channel conditions and the vector of queue sizes $\mathbf{i}$.

The size of the state space of this multidimensional Markov model is $(C+$ $1)^{K} M$ where $C, K$ and $M$ denote the queue size, the number of mobile nodes and the number of different channel conditions. Again, the number of possible transitions is far smaller. There can be arrivals and departures from all queues, as well as state changes of the background process, leading to at most $2 K+M$ transitions.

In contrast to the preceding section, neither sending the arrival rates nor the transmission rates to 0 leads to a triangular generator matrix as changes of the background process can be both upward or downward (whatever the ordering of the state space). However, the generator matrix does have a block triangular representation in both the light-traffic case (arrival rates to 0 ) as well as in the overload case (service rates to 0 ). By this observation, it is possible to calculate the terms in the series expansion around service rate or arrival rate 0 in $O\left(C^{K} M^{3}\right)$, which is only slightly larger than the size of the state space if $M$ is not too large.


Figure 1.7: Queueing model for multi-user wireless access point.

### 1.5 Dissertation outline

This section lays out the structure of the rest of the dissertation. The consecutive chapters focus on the applications introduced in section 1.2 , emerging in the fields of wireless communications and assembly processes. The 4 resulting queueing models are discussed in Chapters 2 to 55. Methodologically, the numerical techniques are based on series expansions and iterative solution methods. Moreover, all numerical results are validated by means of simulation.

Chapter 2 investigates performance of a service-coupled queueing system under intermediate load, motivated by assembly operations, see section 1.2. In particular, the Taylor series expansion of the stationary solution (in the service rate) is calculated around a non-zero service rate. The terms of the expansion are calculated by the weighted Jacobi algorithm. Examples show that the convergence region of the Taylor series expansion allows for a valid approximation in the whole range of intermediate loads by means of a single series expansion.

Chapter 3 introduces an extended version of the model in Chapter 2 with coupled queues, where customers may leave the queue prior to service due to impatience. In the context of the assembly operation, impatience models products with an expiration date. Along with the numerical approximation with Maclaurin series expansions, we also assess the performance of the coupled queueing system by its fluid limit. The fluid approximations complement the Maclaurin series approximations well. Furthermore, a lower bound of the region of convergence of the series expansion is calculated.

Chapter 4 studies the performance evaluation of the multi-user wireless access point that was introduced in Section 1.2. As was already pointed out in Section 1.4 , the light-traffic and overload regimes can be studied efficiently by means of series expansion techniques as the generator matrices have a block triangular structure, the dimensions of the blocks being equal to the number of states of the channel. We show that the model allows for studying an access point with multidimensional Rayleigh fading channels. These channels exhibit both temporal as cross-channel correlation.

Chapter 5 focusses on a queueing model for a Drive-thru Internet scenario in the framework of vehicle-to-infrastructure communication. In order to approximate the performance measures of the roadside base station, we again rely on Taylor series expansions in a wide range of system loads. The terms of the expansion are obtained by the SOR method.

Finally, we draw conclusions in Chapter 6

### 1.6 Publications

The doctoral research performed has resulted in some publications in recognized international research fora, journals and conferences.

### 1.6.1 Publications in international journals

1. E. Evdokimova, S. Wittevrongel, D. Fiems, A Taylor series approach for service-coupled queueing systems with intermediate load, Mathematical Problems in Engineering, 2017, article no. 3298605.

### 1.6.2 Papers in proceedings of international conferences

1. E. Evdokimova, K. De Turck, S. Wittevrongel, D. Fiems, Light-traffic and overload analysis of buffer-aware opportunistic scheduling, Proceedings of the Eighth International Conference on Matrix Analytic Methods in Stochastic Models, MAM8 2014 (Calicut, 8-10 January 2014), pp. 19-21.
2. E. Evdokimova, K. De Turck, S. Wittevrongel, D. Fiems, Efficient performance evaluation of wireless networks with varying channel conditions, Proceedings of the 22nd International Conference on Analytical and Stochastic Modelling Techniques and Applications, ASMTA 2015 (Albena, 26-29 May 2015), Lecture Notes in Computer Science, 2015, vol. 9081, pp. 5972.
3. E. Evdokimova, K. De Turck, S. Wittevrongel, D. Fiems, An analytical performance evaluation tool for wireless access points with opportunistic scheduling, Proceedings of the 9th International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS 2015 (Berlin, 14-16 December 2015), pp. 43-48.
4. E. Evdokimova, S. Wittevrongel, D. Fiems, A Taylor series approach for coupled queueing systems with intermediate load, Proceedings of the 14th International Conference of Numerical Analysis and Applied Mathematics, ICNAAM 2016 (Rhodes, 19-25 September 2016), AIP Conference Proceedings, 2017, vol. 1863, pp. 200003/1-200003/4.

### 1.6.3 Abstracts and other presentations

1. E. Evdokimova, K. De Turck, S. Wittevrongel, D. Fiems, Efficient performance analysis of opportunistic schedulers in wireless networks, Poster presentation at IAP BESTCOM Meeting (Louvain-la-Neuve, 23 October 2014).
2. D. Fiems, E. Evdokimova, K. De Turck, Coupled queues with customer impatience, Book of Abstracts of the Second European Conference on Queueing Theory, ECQT 2016 (Toulouse, 18-20 July 2016), p. 67.

## References

[1] A.K. Erlang. The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik B, 20(33-39):16, 1909.
[2] A. Melikov and L. Ponomarenko. Multidimensional queueing models in telecommunication networks. Springer, 2014.
[3] L. Carrasco, J. Ramis, and G. Femenias. Multidimensional Markov models for the cross-layer design of multi-rate wireless systems using the effective capacity function. In Proceedings of the 69th IEEE Vehicular Technology Conference, pages 2872-2876, 2009.
[4] M. Du, K. Niu, and C. Dong. A Multidimensional Markov Model for Performance Evaluation of Network Selection, Handoff and Traffic Flow Splitting Algorithm in Heterogeneous Wireless Networks. In Proceedings of the 10th International Conference on Communications and Networking in China (CHINACOM), pages 841-846, Shanghai, China, 2015.
[5] Z.T. Lian, L.M. Liu, and M.F. Neuts. A discrete-time model for common lifetime inventory systems. Mathematics of Operations Research, 30(3):718732, 2005.
[6] R. Yang, S. Bhulai, and R. van der Mei. Optimal resource allocation for multiqueue systems with a shared server pool. Queueing Systems, 68(2):133163, 2011.
[7] Oben Ceryan, Izak Duenyas, and Yoram Koren. Optimal control of an assembly system with demand for the end-product and intermediate components. IIE Transactions, 44(5):386-403, 2012.
[8] Q. Xiuhua, C. Chuanhui, and W. Li. A study of some key technologies of $4 G$ system. In Industrial Electronics and Applications, 2008. ICIEA 2008. 3rd IEEE Conference on, pages 2292-2295. IEEE, 2008.
[9] C. Wang, F. Haider, X. Gao, X. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir. Cellular architecture and key technologies for $5 G$ wireless communication networks. IEEE Communications Magazine, 52(2):122-130, 2014.
[10] R. Knopp and P.A. Humblet. Information capacity and power control in single-cell multiuser communications. In IEEE International Conference on Communications, ICC'95 "Gateway to Globalization", volume 1, pages 331335, Seattle, 1995.
[11] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. Scheduling in a queuing system with asynchronously varying service rates. Probability in the Engineering and Informational Sciences, 18(02):191-217, 2004.
[12] J. Lee, R. R Mazumdar, and N.B. Shroff. Opportunistic power scheduling for multi-server wireless systems with minimum performance constraints. In Proceedings of INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, volume 2, pages 10671077, 2004.
[13] A. Asadi and V. Mancuso. A survey on opportunistic scheduling in wireless communications. IEEE Communications Surveys \& Tutorials, 15(4):16711688, 2013.
[14] R. Kwan and C. Leung. A survey of scheduling and interference mitigation in LTE. Journal of Electrical and Computer Engineering, 2010:1, 2010.
[15] J. Ott and D. Kutscher. Drive-thru Internet: IEEE 802.11 b for" automobile" users. In Proceedings of INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, volume 1, 2004.
[16] M. Xing, J. He, and L. Cai. Maximum-utility scheduling for multimedia transmission in drive-thru Internet. IEEE Transactions on Vehicular Technology, 65(4):2649-2658, 2016.
[17] Y. Bozer and L. McGinnis. Kitting versus line stocking: A conceptual framework and a descriptive model. International Journal of Production Economics, 28:1-19, 1992.
[18] P. Som, W.E. Wilhelm, and R.L. Disney. Kitting process in a stochastic assembly system. Queueing Systems, 17(3):471-490, 1994.
[19] H. Brynzér and M.I. Johansson. Design and performance of kitting and order picking systems. International Journal of production economics, 41(1-3):115-125, 1995.
[20] U.N. Bhat. An introduction to queueing theory: modeling and analysis in applications. Birkhäuser, 2015.
[21] D. Beyer, F. Cheng, S.P. Sethi, and M. Taksar. Markovian demand inventory models. Springer, 2010.
[22] G. Liberopoulos, C.T. Papadopoulos, B. Tan, J.M. Smith, and S.B. Gershwin. Stochastic modeling of manufacturing systems. Springer, 2006.
[23] F. Ju, J. Li, and J. A. Horst. Transient analysis of serial production lines with perishable products: Bernoulli reliability model. IEEE Transactions on Automatic Control, 62(2):694-707, 2017.
[24] E. Cinlar. Introduction to stochastic processes. Courier Corporation, 2013.
[25] J.R. Jackson. Networks of waiting lines. Operations research, 5(4):518-521, 1957.
[26] P.G. Harrison. Turning back time in Markovian process algebra. Theoretical Computer Science, 290(3):1947-1986, 2003.
[27] S. Balsamo, G. Dei Rossi, and A. Marin. A Numerical Algorithm for the Solution of Product-Form Models with Infinite State Spaces. In Computer Performance Engineering: 7th European Performance Engineering Workshop, pages 191-206, Bertinoro, Italy, 2010.
[28] H. Chen and D.D. Yao. Fundamentals of queueing networks: Performance, asymptotics, and optimization, volume 46. Springer Science \& Business Media, 2013.
[29] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi. Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. John Wiley \& Sons, 2006.
[30] M.F. Neuts. Matrix-Geometric Solutions to Stochastic Models. Springer, 1984.
[31] Y. Saad. Iterative methods for sparse linear systems. SIAM, 2003.
[32] E. Altman, K.E Avrachenkov, and R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. Advances in Applied Probability, 36(03):839-853, 2004.
[33] J.B. Lasserre. A formula for singular perturbations of Markov chains. Journal of Applied Probability, 31(3):829-833, 1994.
[34] K.E. Avrachenkov, J.A. Filar, and P.G. Howlett. Analytic perturbation theory and its applications. SIAM, 2013.
[35] W.B. van den Hout. The power-series algorithm: a numerical approach to Markov processes. PhD Thesis. Tilburg University, 1996.
[36] G. Koole. On the power series algorithm. CWI, 1994.
[37] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm, pages 53-80. Springer, 1993.
[38] J.P.C. Blanc and R.D. van der Mei. Optimization of polling systems with Bernoulli schedules. Performance Evaluation, 22(2):139-158, 1995.
[39] B. Blaszczyszyn, T. Rolski, and V. Schmidt. Light-traffic approximations in queues and related stochastic models, pages 379-406. CRC Press, 1995.
[40] I. Kovalenko. Rare events in queueing theory. A survey. Queueing systems, 16(1):1-49, 1994.

# A Taylor Series Approach for Service-Coupled Queueing Systems with Intermediate Load 

Ekaterina Evdokimova, Sabine Wittevrongel and Dieter Fiems

published in Hindawi Publishing Corporation Mathematical Problems in Engineering, Mathematical Problems in Engineering, 2017, vol. 2017, Article ID 3298605, 10 pages.


#### Abstract

This paper investigates the performance of a queueing model with multiple finite queues and a single server. Departures from the queues are synchronised or coupled which means that a service completion leads to a departure in every queue and that service is temporarily interrupted whenever any of the queues is empty. We focus on the numerical analysis of this queueing model in a Markovian setting: the arrivals in the different queues constitute Poisson processes and the service times are exponentially distributed. Taking into account the state space explosion problem associated with multidimensional Markov processes, we calculate the terms in the series expansion in the service rate of the stationary distribution of the Markov chain as well as of various performance measures (i) when the system is overloaded and (ii) under intermediate load. Our numerical results reveal that by calculating the series expansions of performance measures around a few service rates, we get accurate estimates of various performance measures once




Figure 2.1: Service-coupled queueing system
the load is above $40 \%$ to $50 \%$.

### 2.1 Introduction

Numerical methods for queueing systems involving multiple queues like queueing networks [1], polling systems [2], priority queues [3] and fork-join queues [4] often suffer from the state-space explosion problem. State-space explosion refers to the problem that multidimensionality of Markov processes leads to processes with a very large state space. Indeed, the size of the state space of a multidimensional Markov process is the product of the number of states in each of its dimensions. Once a few dimensions are involved, the state space becomes very large and direct solution techniques for Markov processes fail. For some particular types of Markov processes, a solution can be readily found, but this depends on structural properties of the Markov chain at hand. We mention Markov chains with product form solutions (like Jackson networks) [5] and M/G/1-type and G/M/1-type Markov processes [6] as particular examples. However many queueing problems do not possess these structural properties, thereby requiring non-standard solution techniques.

This is the case for the queueing system investigated in this paper. We consider a queueing system with $K$ queues in parallel as depicted in Figure 2.1. Customers in all queues receive service simultaneously and there is a departure from every queue upon service completion. Moreover, whenever one of the queues is empty, the server remains idle. That is, an empty queue completely blocks service for all other queues. This queueing system is a natural abstraction for an assembly operation with in-house production. The queues represent inventories for semifinished products which are replenished by in-house production facilities. The final assembly requiring all semi-finished products, the assembly operation is halted once any of the inventories is completely depleted. Finally, the service time of the coupled queueing system represents the assembly time.

We study the service-coupled queueing system under Markovian assumptions. That is, we assume independent Poisson arrivals to all queues with arrival rates
$\lambda_{1}, \ldots, \lambda_{K}$ respectively, and independent exponentially distributed service times with rate $\mu$. Even for these simplified assumptions, the analysis of the coupled queueing system is challenging. First, one cannot impose the often simplifying assumption that queues have infinite capacity as the resulting Markov process is either null recurrent if all arrival rates are equal or transient if this is not the case, see [7] for the coupled queueing system with only two queues. Secondly, the state space of the Markov process for the system with $K$ queues of capacity $C$ is $(C+1)^{K}$ such that a direct solution of the Markov chain is not numerically feasible for moderate $C$ and $K$. Finally, matrix-analytic methods for neither $M / G / 1$-type nor $G / M / 1$-type queueing systems apply, nor is there a product form solution.

To overcome these challenges, literature proposes two alternative approaches, both focusing on approximations for various performance measures of the coupled queueing system. The first approach aims at decomposing the queueing system into a number of independent queueing systems which can be analysed in isolation [8]. Such an analysis approximates the interaction between the different queues by a simpler process which in turn facilitates the analysis. The interaction process is parametrised such that the simplified interaction process corresponds to the expected interaction by the queue in isolation. Alternatively, the system can be studied approximately by means of series expansion techniques if one limits the study to a subset of the parameter space. This is the case in [9, 10] where the coupled queueing system was studied in overload. In these papers it was shown that the terms of the Maclaurin series expansion of the steady-state distribution in the service rate can be obtained at low computational cost. The series expansion of the performance measures can then be easily obtained from the calculated steadystate distribution. However, the numerical approach advocated there only leads to good results when the service rate is close to 0 , or equivalently, when the system is considerably overloaded.

Series expansion techniques for Markov chains go by different names in literature, including perturbation techniques, the power series algorithm and lighttraffic approximations. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of the results with respect to some system parameter. In particular singular perturbations where the perturbation does not preserve the class-structure of the non-perturbed chain, have received considerable attention in literature [11-13]. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by a variable $\rho$. For $\rho=0$, the chain is not only easily solved, but one can also obtain the series expansion in $\rho$. For $\rho=1$ one gets the original Markov chain such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of the series expansion includes $\rho=1$ [14-17]. Finally, light-traffic approximation often corresponds to a series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in
stochastic systems, we further refer the reader to the surveys in [18] and [19].
The present contribution builds on the results of [9] and [10], but considers the service-coupled queueing system when the load of the system is lower. In the context of assembly systems, the overload situation is only natural if assembly is the bottleneck in the production/assembly system. In case production is the actual bottleneck, the assembly queues are not overloaded and the results of [9] and [10] do not apply. However, it is still worth to investigate the assembly system in this case as assembly will be interrupted more often due to a lack of semi-finished products.

Balancing computational cost and accuracy, we investigate the use of Taylor series expansions to calculate the performance measures for a wider range of the service rate. In contrast to the Maclaurin series expansions in [9, 10], the terms in the Taylor series expansion around some service rate $\mu=\mu_{0} \neq 0$ cannot be obtained directly. Therefore we rely on iterative solution methods to solve for the terms in the Taylor series expansion. So, in contrast to the power series algorithm, our approach does not primarily aim for simplifying the solution of the Markov chain, but aims for obtaining the solution in a wide subset of the parameter space at once and relies on iterative procedures to do so.

For any iterative method, a good initial guess of the solution can reduce the number of required iterations considerably. In the present setting, such an initial guess is available if one considers a sequence of Taylor series expansions around increasing values of the service rate starting at $\mu=0$. As shown in [9, 10], the initial series expansion around $\mu=0$ can be calculated efficiently. For higher $\mu$, the expansion around the preceding $\mu$-value can be used to get an initial guess.

The remainder of this paper is organised as follows. The model at hand and the numerical evaluation method are described in the next section. We then illustrate our approach by numerical examples in section 3, prior to drawing conclusions in section 4.

### 2.2 Performance analysis

We consider a queueing system with $K$ finite capacity queues as depicted in Figure 2.1. We denote the capacity of the $k$ th queue by $C_{k}$. The arrival process to the $k$ th queue is assumed to be a Poisson process with a fixed rate $\lambda_{k}$, the arrival processes to the different queues being mutually independent. As mentioned above, service is coupled. This means that there are simultaneous departures from all queues with rate $\mu$ as long as all queues are non-empty, while there are no departures when any of the queues is empty.

In view of the Markovian assumptions on both arrival and service processes, the state (in the Markovian sense) of the queueing system is completely described by the numbers of customers in the different queues. That is, the state of the system
is described by a vector $\mathbf{i}=\left(i_{1}, i_{2}, \ldots, i_{K}\right) \in \mathcal{C}$ where $i_{k}$ denotes the number of customers in the $k$ th queue and where $\mathcal{C}=\left\{0, \ldots, C_{1}\right\} \times \ldots \times\left\{0, \ldots, C_{K}\right\}$ is the state space. We have the following state transitions from state $\mathbf{i} \in \mathcal{C}$ :

- Arrival in queue $k$ (for $k=1, \ldots, K$ ): when $i_{k}<C_{k}$, the arrival rate in queue $k$ is $\lambda_{k}$, the new state being $\mathbf{i}+\mathbf{e}_{k}$. Here $\mathbf{e}_{k}$ is a vector of zeroes, apart from its $k$ th element which is one. There are no arrivals in queue $k$ when $i_{k}=C_{k}$.
- Departure: when all queues are non-empty $\left(i_{1}>0, \ldots, i_{K}>0\right)$ there is a departure from all queues with rate $\mu$. The new state is $\mathbf{i}-\mathbf{e}$, where $\mathbf{e}$ is a vector of ones.

Given the summary of the possible transitions above, the balance equations of the Markov process are readily retrieved. For $\mathbf{i} \in \mathcal{C}$, let $\pi(\mathbf{i})$ be the steady-state probability vector of the queueing system. Equating the total probability flow out of and into state $\mathbf{i}$, we then have the following set of balance equations,

$$
\begin{equation*}
\pi(\mathbf{i})\left(\mu \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\sum_{k=1}^{K} \mathbf{1}_{\left\{i_{k}<C_{k}\right\}} \lambda_{k}\right)=\pi(\mathbf{i}+\mathbf{e}) \mu+\sum_{k=1}^{K} \pi\left(\mathbf{i}-\mathbf{e}_{k}\right) \lambda_{k} \tag{2.1}
\end{equation*}
$$

for $\mathbf{i} \in \mathcal{C}$, where $\mathbf{1}_{\{X\}}$ denotes the indicator function of the event $X$, and where we have assumed $\pi(\mathbf{i})=0$ for $\mathbf{i} \notin \mathcal{C}$ to simplify notation. Since already for a moderate number of queues, the state space is prohibitively large to compute the stationary distribution directly, we rely on a series expansion approach in the remainder.

As the system of equations 2.1 is finite, we find by Cramer's rule that the stationary probabilities $\pi(\mathbf{i})$ can be expressed as rational functions of $\mu$ with at most $M$ distinct poles and no other singularities. Here $M=\prod_{k=1}^{K}\left(C_{k}+1\right)$ is the size of the state space $\mathcal{C}$. Denoting the set of singularities by $\mathcal{M}$, this observation implies that for any $\mu_{0} \in \mathbb{R}^{+} \backslash \mathcal{M}$, the Taylor series expansion in $\mu$ of $\pi(\mathbf{i})$ around $\mu=\mu_{0}$ converges to the correct value in a neighbourhood of $\mu_{0}$. For further reference, let $\pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})$ be the $n$th term in the Taylor series expansion in $\mu$ of $\pi(\mathbf{i})$ around $\mu_{0} \in \mathbb{R}^{+} \backslash \mathcal{M}$. Hence, in a neighbourhood of $\mu_{0}$, we have,

$$
\begin{equation*}
\pi(\mathbf{i})=\sum_{n=0}^{\infty} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} \tag{2.2}
\end{equation*}
$$

First, when $\mu$ is close to 0 , we approximate the stationary probabilities by their Maclaurin series expansion in $\mu$ as investigated in [9]. Plugging the expansion (2.2) for $\mu_{0}=0$ in the balance equations 2.1) and comparing terms in equal
powers of $\mu$, we obtain,

$$
\begin{align*}
\pi_{n}^{(0)}(\mathbf{i}) & \sum_{k=1}^{K} \mathbf{1}_{\left\{i_{k}<C_{k}\right\}} \lambda_{k} \\
& =-\pi_{n-1}^{(0)}(\mathbf{i}) \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\pi_{n-1}^{(0)}(\mathbf{i}+\mathbf{e})+\sum_{k=1}^{K} \pi_{n}^{(0)}\left(\mathbf{i}-\mathbf{e}_{k}\right) \lambda_{k} \tag{2.3}
\end{align*}
$$

for $n \geq 1$ and $\mathbf{i} \neq \mathbf{c}=\left[C_{1}, \ldots, C_{K}\right]$. For $\mathbf{i}=\mathbf{c}$, we find by the normalisation condition,

$$
\begin{equation*}
\pi_{n}^{(0)}(\mathbf{c})=-\sum_{\mathbf{i} \in \mathcal{C} \backslash\{\mathbf{c}\}} \pi_{n}^{(0)}(\mathbf{i}), \tag{2.4}
\end{equation*}
$$

for $n \geq 1$. For $n=0$ and $\mathbf{i} \neq \mathbf{c}$, we further find,

$$
\begin{equation*}
\pi_{0}^{(0)}(\mathbf{i}) \sum_{k=1}^{K} \mathbf{1}_{\left\{i_{k}<C_{k}\right\}} \lambda_{k}=\sum_{k=1}^{K} \pi_{0}^{(0)}\left(\mathbf{i}-\mathbf{e}_{k}\right) \lambda_{k} \tag{2.5}
\end{equation*}
$$

which shows that $\pi_{0}^{(0)}(\mathbf{i})=0$ for $\mathbf{i} \in \mathcal{C} \backslash\{\mathbf{c}\}$ (by evaluation of the expression in lexicographical order). The normalisation condition then further yields $\pi_{0}^{(0)}(\mathbf{c})=$ 1 , such that,

$$
\pi_{0}^{(0)}(\mathbf{i})=\mathbf{1}_{\{\mathbf{i}=\mathbf{c}\}},
$$

for $\mathbf{i} \in \mathcal{C}$. The 0th order terms are trivial and the higher order terms can be calculated one by one in lexicographical order of $\mathbf{i}$ by expressions 2.3) and 2.4) above. The numerical complexity of finding the terms of a single order for all $\mathbf{i} \in \mathcal{C}$ is $O(M K)$ at most. However, one easily verifies that $\pi_{n}^{(0)}(\mathbf{i})=0$ for all i lexicographically smaller than $\mathbf{c}-n \mathbf{e}$, which further reduces the computational complexity of finding the $n$th order terms to $O\left(\min \left(n^{K}, M\right) K\right)$. Note that for large $C_{k}, n^{K}$ is considerably smaller than $M$.

While the terms in the Maclaurin series expansion can be calculated efficiently, the resulting expansion only converges to the exact solution in a neighbourhood of 0 as, in general, the region of convergence of the series expansion will be finite. Therefore, we now consider Taylor series expansions around $\mu=\mu_{0} \neq 0$ to get results for a wider range of the service rate.

Plugging the series expansion (2.2) in the balance equations (2.1) and isolating terms in $\left(\mu-\mu_{0}\right)^{n}$, we get, for $\mathbf{i} \neq \mathbf{c}$,

$$
\begin{align*}
\pi_{0}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu_{0} \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\right. & \left.\sum_{k=1}^{K} \mathbf{1}_{\left\{i_{k}<C_{k}\right\}} \lambda_{k}\right) \\
& =\pi_{0}^{\left(\mu_{0}\right)}(\mathbf{i}+\mathbf{e}) \mu_{0}+\sum_{k=1}^{K} \pi_{0}^{\left(\mu_{0}\right)}\left(\mathbf{i}-\mathbf{e}_{k}\right) \lambda_{k} \tag{2.6}
\end{align*}
$$

and

$$
\begin{align*}
& \pi_{n}^{\left(\mu_{0}\right)(\mathbf{i})\left(\mu_{0} \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\sum_{k=1}^{K} \mathbf{1}_{\left\{i_{k}<C_{k}\right\}} \lambda_{k}\right)} \\
& =-\pi_{n-1}^{\left(\mu_{0}\right)}(\mathbf{i}) \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\pi_{n-1}^{\left(\mu_{0}\right)}(\mathbf{i}+\mathbf{e}) \\
& \quad+\pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i}+\mathbf{e}) \mu_{0}+\sum_{k=1}^{K} \pi_{n}^{\left(\mu_{0}\right)}\left(\mathbf{i}-\mathbf{e}_{k}\right) \lambda_{k} \tag{2.7}
\end{align*}
$$

for $n \geq 1$, whereas the normalisation condition yields,

$$
\begin{equation*}
\sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})=\mathbf{1}_{\{n=0\}} \tag{2.8}
\end{equation*}
$$

In contrast to the Maclaurin expansion above, the system of equations 2.6(2.8) cannot be solved easily. Therefore, we rely on iterative solution methods to find the solution of this system of equations. More specifically, we use weighted Jacobi iteration which calculates the terms in the series expansion by iteratively evaluating,

$$
\begin{aligned}
& \quad \pi_{n, r+1}^{\left(\mu_{0}\right)}(\mathbf{i})=(1-\omega) \pi_{n, r}^{\left(\mu_{0}\right)}(\mathbf{i})+\omega \times \\
& -\pi_{n-1}^{\left(\mu_{0}\right)}(\mathbf{i}) \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\pi_{n-1}^{\left(\mu_{0}\right)}(\mathbf{i}+\mathbf{e})+\pi_{n, r}^{\left(\mu_{0}\right)}(\mathbf{i}+\mathbf{e}) \mu_{0}+\sum_{k=1}^{K} \pi_{n, r}^{\left(\mu_{0}\right)}\left(\mathbf{i}-\mathbf{e}_{k}\right) \lambda_{k} \\
& \mu_{0} \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}+\sum_{k=1}^{K} \mathbf{1}_{\left\{i_{k}<C_{k}\right\}} \lambda_{k}
\end{aligned} .
$$

Here $\omega<1$ denotes the weight of the weighted Jacobi iteration. For each term $n=$ $0,1, \ldots$ and $\mathbf{i} \in \mathcal{C}$, we evaluate for $r=0,1, \ldots$ and approximate $\pi_{n}(\mathbf{i})$ by $\pi_{n, r}(\mathbf{i})$ for $r$ sufficiently large. In practice, we stop iterating when the corresponding terms in the series expansion of the mean and second order moment of the queue content (cf. infra) converge (up to 6 to 8 significant digits).

This iterative approach is computationally feasible as the number of possible transitions from a state is far less than the number of states (the generator matrix is sparse). More precisely, the number of transitions is related to the number of queues such that the numerical complexity of a single iteration for finding the $n$th order terms for all $\mathbf{i} \in \mathcal{C}$ is $O(M K)$.

If $\mu_{0}$ is within the radius of convergence of the preceding expansion, say around $\mu_{0}^{*}$, we use the preceding expansion to get a first approximation for $\pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})$
as to reduce the number of iterations till convergence. That is, we choose,

$$
\begin{aligned}
\pi_{n, 0}^{\left(\mu_{0}\right)}(\mathbf{i}) & =\left.\frac{1}{n!} \frac{\mathrm{d}^{n}}{\mathrm{~d} \mu^{n}}\left(\sum_{m=0}^{N} \pi_{m}^{\left(\mu_{0}^{*}\right)}(\mathbf{i})\left(\mu-\mu_{0}^{*}\right)^{m}\right)\right|_{\mu=\mu_{0}} \\
& =\sum_{m=n}^{N} \pi_{m}^{\left(\mu_{0}^{*}\right)}(\mathbf{i}) \frac{m!}{n!(m-n)!}\left(\mu_{0}-\mu_{0}^{*}\right)^{m-n}
\end{aligned}
$$

If $\mu_{0}$ is not within the radius of convergence of the preceding expansion, we set

$$
\pi_{0,0}^{\left(\mu_{0}\right)}(\mathbf{i})=\prod_{k=1}^{K} \frac{\left(1-\rho_{k}\right) \rho_{k}^{i_{k}}}{1-\rho_{k}^{C_{k}+1}}
$$

with $\rho_{k}=\lambda_{k} /\left(\mu_{0}(1-\alpha)\right)$ and with,

$$
\alpha=1-\prod_{k=1}^{K}\left(1-\frac{1-\lambda_{k} / \mu_{0}}{1-\left(\lambda_{k} / \mu_{0}\right)^{C_{k}+1}}\right) .
$$

That is, we approximate the coupled queueing system, by a queueing system with independent $M / M / 1 / C_{k}$ queues with service rate $\mu_{0}(1-\alpha)$, where $\alpha$ is a crude approximation for the probability that at least one queue is empty. In addition, we set $\pi_{n, 0}^{\left(\mu_{0}\right)}(\mathbf{i})=0$ for $n>0$.

Once the terms in the series expansion are found, we can find approximations for various performance measures. For instance, the $N$ th order expansion of the rth moment of the queue content is calculated as,

$$
\begin{aligned}
\mathrm{E}\left[Q^{\mathbf{r}}\right] \triangleq \mathrm{E}\left[Q_{1}^{r_{1}} Q_{2}^{r_{2}} \ldots Q_{K}^{r_{K}}\right] & \approx \sum_{n=0}^{N} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} \mathbf{i}^{\mathbf{r}} \\
& \triangleq \sum_{n=0}^{N} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} \prod_{k=1}^{K} i_{k}^{r_{k}},
\end{aligned}
$$

where $Q_{k}$ denotes the queue content of the $k$ th queue and with $\mathbf{r}=\left[r_{1}, r_{2}, \ldots, r_{K}\right]$. In particular, the mean $\mathrm{E}\left[Q_{k}\right]$ and variance $\operatorname{var}\left[Q_{k}\right]$ of $Q_{k}$ can be approximated as,

$$
\begin{align*}
& \mathrm{E}\left[Q_{k}\right] \approx \sum_{n=0}^{N} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} i_{k}, \\
& \operatorname{var}\left[Q_{k}\right] \approx \sum_{n=0}^{N} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} i_{k}^{2}-\mathrm{E}\left[Q_{k}\right]^{2} . \tag{2.10}
\end{align*}
$$

Note that the above approximation for the variance is not the $N$ th order series expansion of the variance as the approximation of the square of the mean also
contains terms in $\left(\mu-\mu_{0}\right)^{n}$ for $n>N$. By numerical experimentation, we found that these higher order terms hardly influence the results.

Analogously, let the system content $Q$ be defined as the total number of customers in all queues, then we can approximate the mean $\mathrm{E}[Q]$ and variance $\operatorname{var}[Q]$ of the system content as,

$$
\begin{align*}
& \mathrm{E}[Q] \approx \sum_{n=0}^{N} \sum_{k=1}^{K} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} i_{k}, \\
& \operatorname{var}[Q] \approx \sum_{n=0}^{N} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n}\left(\sum_{k=1}^{K} i_{k}\right)^{2}-\mathrm{E}[Q]^{2} . \tag{2.11}
\end{align*}
$$

Again the same remark applies to the approximation of the variance.
The effective load is defined as the fraction of time that the server is serving. As the server is serving whenever all queues are non-empty, we find the following $N$ th order expansion of the effective load $\rho_{\text {eff }}$,

$$
\rho_{\mathrm{eff}} \approx \sum_{n=0}^{N} \sum_{\mathbf{i} \in \mathcal{C}} \pi_{n}^{\left(\mu_{0}\right)}(\mathbf{i})\left(\mu-\mu_{0}\right)^{n} \prod_{k=1}^{K} \mathbf{1}_{\left\{i_{k}>0\right\}}
$$

Finally, let the blocking probability be the fraction of customers that cannot be accepted upon arrival in the queueing system. The effective load allows for calculating the blocking probability $b_{k}$ in the $k$ th queue. Indeed, noting that all accepted customers must be served, we have,

$$
\lambda_{k}\left(1-b_{k}\right) \frac{1}{\mu}=\rho_{\mathrm{eff}}
$$

or, equivalently,

$$
b_{k}=1-\rho_{\mathrm{eff}} \frac{\mu}{\lambda_{k}}
$$

Notice that $b_{k}$ only depends on the queue capacity through $\rho_{\text {eff }}$. The latter is influenced by the capacities of all the different queues, which particularly implies that the capacity of one queue influences the blocking probabilities of the other queues.

### 2.3 Numerical results

We now evaluate our numerical approximation approach by some numerical examples. We focus on the mean and standard deviation of the queue content as well as on the blocking probability. Noting that in a coupled queueing system with non-equal arrival loads, the performance is mainly determined by the queues with the lowest loads (the queues with higher load can be neglected when studying the overall performance), we first focus on a coupled queueing system with an equal


Figure 2.2: Nth order approximations for heavy and intermediate traffic for the mean queue content of the coupled queueing system with $K=5$ queues, each having capacity $C=15$ and arrival rate $\lambda=1$ for each queue.
arrival rate $\lambda$ in all queues. Without loss of generality, we set $\lambda=1$ (as we can scale $\mu$ to investigate a different $\lambda$ ). We consider $K=5$ queues, each having capacity $C=15$.

Figures $2.2,2.3$ and 2.4 depict the mean queue content versus the service rate $\mu$, the blocking probability versus the service rate $\mu$, and the standard deviation of the queue content versus $\mu$, respectively. Note that we have the same blocking probability and the same mean and variance of the queue content for every queue due to symmetry and that we approximate the standard deviation of the queue content by $\sqrt{\operatorname{var}\left[Q_{k}\right]}$ with $\operatorname{var}\left[Q_{k}\right]$ given in 2.10. Each figure shows the 5 th, 15 th and 30th order approximation on a separate subfigure, and we combine the Maclaurin expansion around 0 with the approximation around $\mu_{0}=1.5$ for all performance measures. For visual reference, the point $\mu_{0}$ is marked on all the figures with a cross. The order $N$ of the expansion refers to both the order of the expansion around 0 and the order of the expansion around $\mu_{0}$. In addition, we show simulation results for the performance measures at hand, which allows for evaluating the accuracy of the approximations. We used uniformisation to simulate the queueing system (based on the balance equations) and generated $10^{8}$ samples, for each simulation point. We calculated the confidence interval by means of the batch means method, but omitted the confidence intervals from the plots as the obtained upper and lower bounds are visually indiscernible.

For the coupled queueing system under study with $K=5$ queues of capacity $C=15$, the Markov chain has $M=1.048 .576$ states. The figures show that the approximations of the mean queue content and the blocking probability are already fairly accurate for the 5 th order expansion $(N=5)$, whereas the standard deviation


Figure 2.3: Nth order approximations for heavy and intermediate traffic for the blocking probability of the coupled queueing system with $K=5$ queues, each having capacity $C=15$ and arrival rate $\lambda=1$ for each queue.


Figure 2.4: Nth order approximations for heavy and intermediate traffic for the standard deviation of the queue content of the coupled queueing system with $K=5$ queues, each having capacity $C=15$ and arrival rate $\lambda=1$ for each queue.


Figure 2.5: Nth order approximations for heavy and intermediate traffic for the mean queue content for two asymmetric queues of the coupled queueing system with $K=6$ queues, each having capacity $C=10$, and arrival rates $\lambda_{1}=1$ for half of the queues and $\lambda_{2}=2$ for the rest.
of the queue content requires some more terms ( $N=15$ ). As the order $N$ of the expansions further increases, the approximations even more closely approximate the performance measures at hand. The figures further reveal that the match is very good in a limited region (of 0 or of $\mu_{0}$ ), while the approximations quickly grow to very large values outside this region. This is not unexpected as the region of convergence is finite for sure ( $\pi(\mathbf{i})$ is a rational function of $\mu$, cf. supra). While the sharp deterioration of the approximation prevents one to extend the results outside the region of convergence of the series expansion, it does give a clear indication where the approximation is accurate. Overall, we find that the 30th order approximations for the mean queue content, the blocking probability and the standard deviation are accurate for loads above $45 \%$ ( $\mu$ below 2.25).

The effect of increasing $\mu$ on the mean queue content and on the blocking probability confirms intuition. If the service speed increases, the mean content decreases and as it is less likely that the queues are full, the blocking probability decreases as well. The decrease is fast for low $\mu$ and slower for larger $\mu$, the change of the decay rate being around $\mu=1$ (or a load of $100 \%$ ) for the blocking probability and just above $\mu=1$ for the mean queue content. For the standard deviation, we observe that it increases with $\mu$.

Next, we study an example with non-equal arrival rates at the different queues. In particular, we consider a system with $K=6$ queues, each having capacity $C=10$, which results in a Markov chain with $M=1.771 .561$ states. In order to investigate the impact of non-equal arrival loads, we consider a system with two arrival rates: arrival rate $\lambda_{1}=1$ for half of the queues and arrival rate $\lambda_{2}=2$ for


Figure 2.6: Nth order approximations for heavy and intermediate traffic for the blocking probability for two asymmetric queues of the coupled queueing system with $K=6$ queues, each having capacity $C=10$, and arrival rates $\lambda_{1}=1$ for half of the queues and $\lambda_{2}=2$ for the rest.


Figure 2.7: Nth order approximations for heavy and intermediate traffic for the standard deviation of the queue content for two asymmetric queues of the coupled queueing system with $K=6$ queues, each having capacity $C=10$, and arrival rates $\lambda_{1}=1$ for half of the queues and $\lambda_{2}=2$ for the rest.


Figure 2.8: 30 th order approximations for heavy and intermediate traffic for the mean queue content and blocking probability of three queueing systems with 2, 4 and 6 queues; for each system queue capacity $C=10$, arrival rates $\lambda=1$.
the remaining queues.
Figures 2.5, 2.6 and 2.7 depict the mean queue content, the blocking probability, and the standard deviation of the queue content versus the service rate $\mu$, respectively, for queues with arrival rate $\lambda_{1}$ as well as for queues with arrival rate $\lambda_{2}$. We again depict approximations of order $N=5,15$ and 30 on different subfigures. For every order $N$, we consider the expansion around 0 and the expansion around $\mu_{0}=1.5$, the point $\mu_{0}$ being marked with a cross on all plots. The plots again reveal that the approximations are quite accurate, especially the 30 th order approximation which is again accurate for $\mu$ up to 2.25 . An increase of $\mu$ leads to a decrease of the mean queue content and of the blocking probability as for the symmetric case, while it leads to an increase of the standard deviation of the queue content. Also, the queues with the highest arrival rate $\left(\lambda_{2}\right)$ have higher mean queue content and blocking probability as there are more arrivals, which also leads to a reduction of the standard deviation of the queue content, as the more heavily loaded queue is close to full most of the time.

As a final example, we assess the impact of the number of queues involved. To this end, we compare the performance of the queueing system with $K=2, K=4$ and $K=6$ queues. All queues have capacity $C=10$ and equal arrival rate $\lambda=1$. Figure 2.8 shows the 30th order approximations (in 0 and 1.5) for the mean queue content and the blocking probability as a function of the service rate $\mu$. We can readily observe that adding queues leads to performance degradation (higher mean queue content and higher blocking probability), especially when the system is not in overload. This is not unexpected as it is more likely that one of the queues is empty in systems with more queues. For coupled queueing systems in overload,
the number of queues involved has hardly any impact on performance though. In overload, it is unlikely that queues are empty, so the number of queues does not matter.

### 2.4 Conclusions

In this paper we presented a numerical approach for the performance evaluation of coupled queueing systems. The study was motivated by an assembly-like system, where inventory replenishments can be modelled by Poisson processes. The presented method focuses on coupled queueing systems working under intermediate load and builds on a previously designed method for such systems in overload. We showed that the region where an accurate estimation is obtained, can be extended to lower loads by iteratively calculating the terms of the Taylor series expansion of the steady-state probability vector.

An important contribution of the study is that the problem is tackled numerically, while existing analysis methods for large-scale queueing systems mainly rely on simulation. We showed that our analysis method allows for performance evaluation under intermediate load, although the specific region of accuracy may vary depending on the system size and structure.

## References

[1] Y.V. Malinkovskii. Jackson networks with single-line nodes and limited sojourn or waiting times. Automation and Remote Control, 76(4):603-612, 2015.
[2] K. Avrachenkov, E. Perel, U. Yechiali. Finite-buffer polling systems with threshold-based switching policy. TOP, 24(3):541-571, 2016.
[3] T. Maertens, J. Walraevens, H. Bruneel. Priority queueing systems: from probability generating functions to tail probabilities. Queueing Systems, 55(1):2739, 2007.
[4] A. Thomasian. Analysis of fork/join and related queueing systems. ACM Computing Surveys, 47(2):17, 2015.
[5] W. Henderson, P.G. Taylor. Product form in networks of queues with batch arrivals and batch services. Queueing Systems, 6(1):71-87, 1990.
[6] D.A. Bini, G. Latouche, B. Meini. Numerical methods for structured Markov chains. Oxford University Press, 2005.
[7] G. Latouche. Queues with paired customers. Journal of Applied Probability 18(3):684-696, 1981.
[8] W.J. Hopp, J.T. Simon. Bounds and heuristics for assembly-like queues. Queueing Systems, 4(2):137-155, 1989.
[9] K. De Turck, E. De Cuypere, S. Wittevrongel, D. Fiems. Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems. In Proceedings of the 6th International Conference on Performance Evaluation Methodologies and Tools, pages 38-44, 2012.
[10] E. De Cuypere, K. De Turck, D. Fiems. A Maclaurin-series expansion approach to multiple paired queues. Operations Research Letters, 42(3):203207, 2014.
[11] J.B. Lasserre. A formula for singular perturbations of Markov chains. Journal of Applied Probability, 31(3):829-833, 1994.
[12] E. Altman, K.E. Avrachenkov, R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. Advances in Applied Probability, 36(3):839-853, 2004.
[13] K.E. Avrachenkov, J.A. Filar, P.G. Howlett. Analytic perturbation theory and its applications. SIAM, 2013.
[14] W.B. van den Hout. The power-series algorithm: a numerical approach to Markov processes. PhD Thesis, Tilburg University, 1996.
[15] G. Koole. On the power series algorithm. CWI, 1994.
[16] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In Performance Evaluation of Computer and Communication Systems, pages 53-80, 1993.
[17] J.P.C. Blanc, R.D. van der Mei. Optimization of polling systems with Bernoulli schedules. Performance Evaluation, 22(2):139-158, 1995.
[18] B. Błaszczyszyn, T. Rolski, V. Schmidt. Light-traffic approximation in queues and related stochastic models. In Advances in Queueing: Theory, Methods, and Open Problems, pages 379-406, 1995.
[19] I.N. Kovalenko. Rare events in queueing systems - A survey. Queueing Systems, 16(1):1-49, 1994.

# Coupled queues with customer impatience 

Ekaterina Evdokimova, Koen De Turck and Dieter Fiems

submitted to Performance Evaluation.


#### Abstract

Motivated by assembly processes, we consider a Markovian queueing system with multiple coupled queues and customer impatience. Coupling means that departures from all constituent queues are synchronised and that service is interrupted whenever any of the queues is empty and only resumes when all queues are non-empty again. Even under Markovian assumptions, the state-space grows exponentially with the number of queues involved. To cope with this inherent state-space explosion problem, we investigate performance by means of two numerical approximation techniques based on series expansions, as well as by deriving the fluid limit. In addition, we provide closed form expressions for the first terms up in the series expansion of the mean queue content for the symmetric coupled queueing system. By means of an extensive set of numerical experiments, we show that the approximation methods complement each other, each one being accurate in a particular subset of the parameter space.


### 3.1 Introduction

We investigate the performance of a particular Markovian queueing system with $K$ parallel queues, as depicted in Figure 3.1. The queues have finite or capacity; let $C_{k} \in \mathbb{N}^{+}$be the capacity of the $k$ th queue. Customers arrive at the $k$ th queue in accordance with a Poisson process with rate $\lambda_{k}>0$, the arrival processes at the different queues being independent. We further assume that departures from the different queues are coupled. This means that there are simultaneous departures from all queues with rate $\mu$ as long as all queues are non-empty. If one of the queues is empty, no service takes place. Finally, customer impatience is assumed: each customer leaves the $k$ th queue prior to service with abandonment rate $\alpha_{k}$ with the exception of customers whose service has started.

The queueing system described above is a natural abstraction for an assembly process with multiple inventories; see [1, 2] and the references therein for advances in stochastic inventory models. The different queues represent part inventories for the different parts that are used during assembly. These inventories are continuously replenished by in-house production facilities (in accordance to a Poisson process), the inventories offering temporary storage to smooth out uncertainty in the various production processes. Parts are assumed to be perishable, meaning that they should be used before a (random) due-date or be discarded once this due-date is crossed. This perishability is captured by the abandonment processes from the different queues. Food-products are a prime example of perishable semifinished products. However, perishable semi-finished products are also found in biochemical production, and in battery and semiconductor manufacturing [3]. Finally, assuming that assembly requires that all the necessary inputs are available, it can only proceed if the inventories (or queues) are not empty, which corresponds to the notion of the coupled departures introduced above.

The two-buffer coupled queueing system without customer impatience is well understood. If the buffer capacity is infinite, the uncontrolled queue process is null recurrent in the Markovian setting. The inherent instability of such queueing systems is demonstrated in [6] where the buffer content difference is studied in the two-queue case. Assuming finite capacity buffers, Hopp and Simon developed a model for a two-buffer kitting process with exponentially distributed processing times for kits and Poisson arrivals [5]. The exponential service times and Poisson arrival assumptions were later relaxed in [16] and [17], respectively.

Only a few authors have studied coupled (or paired) queueing systems with multiple (i.e. more than two) queues. In [4], Harrison studies stability of coupled queueing under very general assumptions: $K \geq 2$ infinite-capacity buffers, generally distributed interarrival times at the different buffers, and generally distributed service times. He proves that stability requires buffer control, or more precisely, that the distribution of the vector of waiting times (in the different queues) with-


Figure 3.1: Representation of the coupled queueing system with customer impatience
out control and infinite queue capacity is defective. When the queues are finite, such a control is not necessary. The queue content of the coupled queueing system with finite buffers is studied in [18], assuming exponential service and Poisson arrivals. As the size of the state-space of the associated Markov chain grows quickly with the number of queues involved, [18] presents an approximation for the queue content when the system is in overload.

In contrast to the uncontrolled coupled queueing system, the controlled coupled queueing system has received considerable attention in the scientific literature. Ramakrishnan and Krishnamurthy adopt the term synchronisation station and present a recent account on approximations of such systems [8]. A particular type of control of coupled queues relates to fork-join type queueing system [9, 10]. In fork-join systems, a job is forked into different sub-jobs, run on different servers. Upon completion of all sub-jobs, there is a final service joining the sub-jobs again. The server joining the sub-jobs operates as a coupled server, albeit with a controlled arrival process. Indeed, the sub-jobs that need to be merged, are already present in the fork-join system. These will be available for the coupled server after some delay.

Coupled queueing may also refer to different types of multi-queueing systems, most prominently to systems with discriminatory processor sharing. In discriminatory processor sharing the total service capacity is distributed amongst all queues that have waiting customers, some queues getting a larger share than others. Once one of the queues is empty, its share is moved to the queues with waiting customers. The authors in [11] investigate such a two-queue system where customers are served in both queues at unit rate when both queues are non-empty, while the non-empty queue is served at a higher rate when the other is empty. A similar system is studied in [12] in the heavy traffic regime while [13] allows for time varying arrival rates and the possibility of jobs abandoning. In contrast to [11-13],
jobs in the first queue do not leave the system but move to the second queue upon completion in [14]. Finally, [15] studies the stability of a more generic system with multiple queues where the service rate of each queue depends on the number of customers in all queues.

The present paper investigates approximations for multi-buffer coupled queuing systems with customer impatience. We investigate two numerical approximation techniques as well as the fluid limit of the system at hand. The numerical approximation methods rely on a Maclaurin-series expansion of the steady-state probability vector, either around $\lambda=0$ (light-traffic) or around $\alpha=\mu=0$ (overload). Series expansion techniques for Markov chains are referred to as perturbation techniques, the power series algorithm and light-traffic approximations. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of performance measures with respect to the system parameters. In particular singular perturbations where the perturbation does not preserve the class-structure of the non-perturbed chain, have received considerable attention in literature, see [19-21] and the references therein. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by an auxiliary variable $\rho$. For $\rho=0$, the chain can be solved efficiently, and one can also obtain the perturbation of the chain in $\rho$. For $\rho=1$ the original Markov chain is retrieved such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of the series expansion includes $\rho=1$, see e.g. [22-25]. Finally, light-traffic approximations often corresponds to a series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in stochastic systems, we further refer the reader to the surveys in [26] and [27].

The remainder of the paper is organised as follows. In the next section, we introduce the balance equations and present the numerical light-traffic analysis. Performance in overload is investigated in section 3.3 while section 3.4 focuses on the fluid limit when $\alpha_{n}>0$ and $\mu<\lambda_{n}$. Finally, we assess the accuracy of the approximations by means of numerical examples in section 3.5 and draw conclusions in section 3.6

### 3.2 Light traffic analysis

We first derive the balance equations for the coupled queueing system. In view of the modelling assumptions introduced above, the state of the coupled queueing system is described by the number of customers in the queues. Let $X_{k}(t)$ be the number of customers in the $k$ th queue at time $t$ and let $\mathbf{X}(t)=\left[X_{1}(t), \ldots, X_{K}(t)\right] \in$ $\mathcal{X}$, where $\mathcal{X}$ denotes the state space of the Markov chain,

$$
\mathcal{X}=\left\{0,1, \ldots, C_{1}\right\} \times \ldots \times\left\{0,1, \ldots, C_{K}\right\} .
$$

Further, let $\pi(\mathbf{x})=\lim _{t \rightarrow \infty} \mathrm{P}[\mathbf{X}(t)=\mathbf{x}]$ be the stationary probability vector of the process, for $\mathbf{x}=\left[x_{1}, \ldots, x_{K}\right] \in \mathcal{X}$. In particular, $\pi(\mathbf{x})=0$ for $\mathbf{x} \notin \mathcal{X}$ which simplifies the notation.

The following notation is introduced for further use. Let $1_{\{\cdot\}}$ be the indicator function which evaluates to one if its argument is true and to 0 if its argument is false. The vector $\mathbf{e}_{k}=\left[1_{\{\ell=k\}}\right]_{\ell=1, \ldots, K}$ denotes a row vector with all its elements zero, apart from the $k$ th element which is 1 , whereas $\mathbf{e}=\sum_{k} \mathbf{e}_{k}$ denotes a row vector with $K$ ones. Given the description of the queueing system and its notation in section 3.1, and the notation introduced above, we can now summarise the possible state transitions from state $\mathbf{x}$.

- Provided that the $k$ th queue is not full $\left(x_{k}<C_{k}\right)$, new customers arrive at this queue with rate $\lambda_{k}$, inducing a transition to state $\mathbf{x}+\mathbf{e}_{k}$.
- Provided that no queue is empty ( $x_{k}>0$ for all $k$ ), there is a departure event with rate $\mu$. A departure event corresponding to a single departure from each queue results in the new state being $\mathbf{x}-\mathbf{e}$.
- Finally, customers abandon the $k$ th queue with rate $\alpha\left(x_{k}-1\right)$ if all queues are non-empty (as the customer being served does not abandon) and with rate $\alpha x_{k}$ if at least one queue is empty (in this case, there is no service). After the abandonment in the $k$ th queue, the new state is $\mathbf{x}-\mathbf{e}_{k}$.

Accounting for the different types of transitions, we find the following set of balance equations,
$\pi(\mathbf{x}) A(\mathbf{x})=\pi(\mathbf{x}+\mathbf{e}) \mu+\sum_{k=1}^{K} \pi\left(\mathbf{x}-\mathbf{e}_{k}\right) \lambda_{k}+\sum_{k=1}^{K} \pi\left(\mathbf{x}+\mathbf{e}_{k}\right) \alpha_{k}\left(x_{k}+1-E\left(\mathbf{x}+\mathbf{e}_{k}\right)\right)$
for $\mathrm{x} \in \mathcal{X}$, with,

$$
A(\mathbf{x})=\sum_{k=1}^{K} \alpha_{k}\left(x_{k}-E(\mathbf{x})\right)+\sum_{k=1}^{K} \lambda_{k} 1_{\left\{x_{k}<C_{k}\right\}}+\mu E(\mathbf{x})
$$

and with $E(\mathbf{x})$ the indicator function that all queues are non-empty,

$$
E(\mathbf{x})=\prod_{k=1}^{K} 1_{\left\{x_{k}>0\right\}}
$$

For the light-traffic approximation, we express all $\lambda_{k} \doteq \kappa_{k} \lambda$ in terms of $\lambda$ and we then send $\lambda$ to zero. The system of balance equations (3.1) has a matrix representation

$$
\begin{equation*}
\boldsymbol{\pi} \mathcal{A}=\boldsymbol{\pi}\left(\mathcal{A}_{0}+\lambda \mathcal{A}_{1}\right)=0 \tag{3.2}
\end{equation*}
$$

where $\boldsymbol{\pi}=[\pi(\mathbf{x})]_{x \in \mathcal{X}}$ is the stationary probability vector, and where $\mathcal{A}, \mathcal{A}_{0}$ and $\mathcal{A}_{1}$ are $S \times S$ matrices that do not depend on $\lambda$. Here $S=|\mathcal{X}|$ denotes the size of the state space,

$$
\begin{equation*}
S=\prod_{k=1}^{K}\left(C_{k}+1\right) \tag{3.3}
\end{equation*}
$$

Note that $\mathcal{A}_{0}$ only contains transition rates corresponding to service completions and/or abandonments, while $\mathcal{A}_{1}$ only contains transitions corresponding to arrivals.

### 3.2.1 Numerical series expansion

Direct solution of the system of equations (3.1), or of (3.2), is only possible if the number of queues and their capacities is limited, as the size of the state space grows quickly with the number of queues, see (3.3). Therefore, we introduce the Maclaurin series expansion of the stationary probability $\pi(\mathbf{x})$,

$$
\pi(\mathbf{x})=\sum_{n=0}^{\infty} \pi_{n}(\mathbf{x}) \lambda^{n}
$$

or, equivalently, of the stationary vector $\pi$,

$$
\begin{equation*}
\boldsymbol{\pi}=\sum_{n=0}^{\infty} \boldsymbol{\pi}_{n} \lambda^{n} \tag{3.4}
\end{equation*}
$$

This series expansion is justified in section 3.2.3 where a lower bound for the region of convergence of the series expansion is calculated.

Plugging (3.4) into (3.2) and equating the terms in $\lambda^{n}$ yields,

$$
\begin{equation*}
\pi_{0} \mathcal{A}_{0}=0, \quad \boldsymbol{\pi}_{n} \mathcal{A}_{0}=-\boldsymbol{\pi}_{n-1} \mathcal{A}_{1} \tag{3.5}
\end{equation*}
$$

for $n \in \mathbb{N}^{+}$, whereas the normalisation condition $\pi \mathbf{e}^{\prime}=1$ yields,

$$
\boldsymbol{\pi}_{0} \mathbf{e}^{\prime}=1, \quad \boldsymbol{\pi}_{n} \mathbf{e}^{\prime}=0
$$

for $n \in \mathbb{N}^{+}$.
Assuming that the states are ordered lexicographically, one finds that $\mathcal{A}_{0}$ is lower triangular as $\mathcal{A}_{0}$ collects the transition rates corresponding to departures (either by impatience or after a service completion). As a consequence, the recursive equations (3.5) can be readily solved. We express the recursion in terms of the system parameters below.

In absence of arrivals $(\lambda=0)$, the stationary solution is the empty queue. That is, $\boldsymbol{\pi}_{0}$ equals,

$$
\pi_{0}(\mathbf{x})= \begin{cases}1 & \text { for } x_{1}=0, \ldots, x_{K}=0 \\ 0 & \text { otherwise }\end{cases}
$$

Given $\pi_{0}$, we can calculate the higher order terms recursively. Given the $(n-1)$ st vector $\boldsymbol{\pi}_{n-1}$, we can calculate the values $\pi_{n}(\mathbf{x})$ in reverse lexicographical order by,

$$
\begin{gather*}
\pi_{n}(\mathbf{x})=\left(\sum_{k=1}^{K} \pi_{n-1}\left(\mathbf{x}-\mathbf{e}_{k}\right) \kappa_{k}-\pi_{n-1}(\mathbf{x}) \sum_{i=1}^{K} \kappa_{i} 1_{\left\{x_{i}<C_{i}\right\}}+\pi_{n}(\mathbf{x}+\mathbf{e}) \mu\right. \\
\left.+\sum_{k=1}^{K} \pi_{n}\left(\mathbf{x}+\mathbf{e}_{k}\right) \alpha_{k}\left(x_{k}+1-E\left(\mathbf{x}+\mathbf{e}_{k}\right)\right)\right) \\
\times\left(\sum_{k=1}^{K} \alpha_{k}\left(x_{k}-E(\mathbf{x})\right)+\mu E(\mathbf{x})\right)^{-1} \tag{3.6}
\end{gather*}
$$

for $\mathbf{x} \neq[0,0, \ldots, 0] \doteq \mathbf{0}$. Finally, for $\mathbf{x}=\mathbf{0}$, the normalisation condition yields,

$$
\pi_{n}(\mathbf{0})=-\sum_{\mathbf{x} \in \mathcal{X} \backslash\{\mathbf{0}\}} \pi_{n}(\mathbf{x})
$$

Remark 1. The recursion above closely resembles the well-known Gauss-Seidel method. Indeed, the transition matrix $\mathcal{A}$ is decomposed into a lower and upper triangular matrix, which yields a recursion where each step is easily solved. In the present setting, the Gauss-Seidel method allows for calculating the stationary probability vector for a single value $\lambda$. In contrast, we obtain a polynomial expression for the stationary probability vector which accurately approximates the probability vector in some interval [ $0, \lambda_{\text {max }}$ ].
Remark 2. The above recursion can be used when the capacity of some (or all) of the queues is infinite. Indeed, let $\mathcal{X}_{n}=\{\mathbf{x} \in \mathcal{X},|\mathbf{x}| \leq n\}$ be the set of system states where the total system content does not exceed $n$. A careful analysis of the recursion above reveals that $\pi_{n}(\mathbf{x})=0$ for $\mathbf{x} \in \mathcal{X} \backslash \mathcal{X}_{n}$. Hence, the number of non-zero terms in the $n$th order expansion is finite, even if the queue capacity is infinite. This observation confirms the so-called $n$-events rule which states that for an $n$th order expansion, only sample paths with $n$ or fewer perturbed events must be considered [26].
Remark 3. When the capacity of all queues is infinite, the complexity of calculating $N$ terms in the expansion is $O\left(N^{K+1} K\right)$. In view of the preceding remark, the number of non-zero values in $\pi_{n}$ is $O\left(n^{K}\right)$, the calculation of a single term having complexity $O(K)$. When the buffer size is finite, the number of values to calculate is also bounded by the size of the state space. Assuming buffers with equal finite capacity $C$, the computational complexity of calculating $N$ terms in the expansion is $O\left(\min (C, N)^{K} K N\right)$. Indeed, for each term in the series expansion, we need to calculate $(C+1)^{K}$ values at most.
Remark 4. The computational complexity further decreases when the arrival rates and abandonment rates in the different queues are equal. By symmetry, one then
has $\pi_{n}(\mathbf{x})=\pi_{n}(\mathbf{y})$ for any permutation $\mathbf{y}$ of $\mathbf{x}$. Limiting the discussion to the case of infinite capacity buffers (which naturally forms an upper bound), the number of values to calculate for the $n$th order term is [29]

$$
c_{n}=\sum_{m=0}^{n} p_{K}(m+K)
$$

Here $p_{k}(n)$ is the number of partitions of the integer $n$ into exactly $k$ positive integer parts, satisfying the recursion,

$$
p_{k}(n)=p_{k}(n-k)+p_{k-1}(n-1), \quad p_{0}(0)=1
$$

assuming $p_{k}(n)=0$ for $k>n$. The first 10 values of the sequence $c_{n}$ for any $C>10$ are given below,

$$
1,2,4,7,12,19,30,45,67,97, \ldots
$$

### 3.2.2 Closed form expressions for the symmetric coupled queueing system

For the symmetric coupled queueing system we obtain closed-form expressions for the $K$ th order expansion of the first two moments of the queue content. As the system is symmetric we have $\alpha_{k}=\alpha$ and $\kappa_{k}=1$ for $k=1, \ldots, K$. In addition, we assume that the queue capacities exceed $K: C_{k}>K$ for all $k=1, \ldots, K$.

Repeated application of the set of recursive equations, then yields the following series expansions of the first two moments of the queue content $X$ (that is, the content of an arbitrary queue),

$$
\begin{align*}
\mathrm{E}[X] & =\frac{1}{\alpha} \lambda-\frac{K(\mu-\alpha)}{\mu \alpha^{K}} \lambda^{K}+O\left(\lambda^{K+1}\right),  \tag{3.7}\\
\mathrm{E}\left[X^{2}\right] & =\frac{1}{\alpha} \lambda+\frac{1}{\alpha^{2}} \lambda^{2}-\frac{K(\mu-\alpha)}{\mu \alpha^{K}} \lambda^{K}+O\left(\lambda^{K+1}\right) \tag{3.8}
\end{align*}
$$

Notice the disappearing terms in the power expansion (from 2 up to $K-1$ in case of $\mathrm{E}[X]$, from 3 to $K-1$ in case of $\mathrm{E}\left[X^{2}\right]$ ). This can be explained by the $n$ events rule: for the $n$th order expansion in $\lambda$ we need to consider only $n$ arrivals, and when $n<K$, there are only departures due to impatience (and not due to service completion), hence it can be intuited that the first term containing the parameter $\mu$ is indeed of the $K$ th order.

### 3.2.3 Lower bound for the radius of convergence

We now focus on a lower bound for the radius of the series expansion. The basic ideas for finding such a bound date back to the seminal work of Schweitzer [28].

We validate the series expansion by explicitly constructing the expansion. To do so, we first introduce some additional notation and the basic notion of the deviation matrix of a CTMC.

Let $\boldsymbol{\pi}^{(\lambda)}$ denote the steady state solution $[\pi(\mathbf{x})]_{\mathbf{x} \in \mathcal{X}}$ of the balance equations (3.1). We have made the dependence of $\boldsymbol{\pi}^{(\lambda)}$ on $\lambda$ explicit for ease of notation. With this notation, the balance equations can be written in matrix notation as follows,

$$
\begin{equation*}
\boldsymbol{\pi}^{(\lambda)} \mathcal{A}^{(\lambda)}=\boldsymbol{\pi}^{(\lambda)}\left(\mathcal{A}_{0}+\lambda \mathcal{A}_{1}\right)=0 \tag{3.9}
\end{equation*}
$$

see equation 3.2. In view of the system assumptions it is readily seen that $\mathcal{A}^{(0)}=\mathcal{A}_{0}$ only has one recurrent state, i.e. $\mathbf{0}$ (the empty state) is recurrent and all the others are transient. Therefore, the stationary vector $\boldsymbol{\pi}^{(0)}$ exists, with state $\pi^{(0)}(\mathbf{0})=1$ and $\pi^{(0)}(\mathbf{x})=0$ for $\mathbf{x} \in \mathcal{X} \backslash\{\mathbf{0}\}$.

Let $\mathcal{D}_{0}$ be the deviation matrix of the CTMC with generator matrix $\mathcal{A}_{0}$,

$$
\begin{equation*}
\mathcal{A}_{0}=\int_{0}^{\infty}\left(\mathcal{P}_{0}(t)-\boldsymbol{\Pi}_{0}\right) d t \tag{3.10}
\end{equation*}
$$

Here the family $\left\{\mathcal{P}_{0}(t)=\exp \left(\mathcal{A}_{0} t\right), t \geq 0\right\}$ is the Markov semigroup of the CTMC, and $\Pi_{0}=\lim _{t \rightarrow \infty} \mathcal{P}_{0}(t)=\mathbf{e}^{\prime} \boldsymbol{\pi}^{(0)}$, $\mathbf{e}^{\prime}$ being a column vector of ones. As the state-space $\mathcal{X}$ is finite, the deviation matrix is well defined. Moreover, the deviation matrix satisfies $\mathcal{D}_{0} \mathbf{e}^{\prime}=0$ - the row sums are zero - and,

$$
\begin{equation*}
\mathcal{D}_{0} \mathcal{A}_{0}=\mathcal{A}_{0} \mathcal{D}_{0}=\boldsymbol{\Pi}_{0}-\mathcal{I} \tag{3.11}
\end{equation*}
$$

with $\mathcal{I}$ the identity matrix.
Theorem 2. The solution $\boldsymbol{\pi}^{(\lambda)}$ of the CTMC adheres to the following power series expansion,

$$
\begin{equation*}
\boldsymbol{\pi}^{(\lambda)}=\sum_{k=0}^{\infty}\left(\boldsymbol{\pi}^{(0)}\left(\mathcal{A}_{1} \mathcal{D}_{0}\right)^{k}\right) \lambda^{k} \tag{3.12}
\end{equation*}
$$

for $0 \leq \lambda<\lambda_{0}, \lambda_{0}^{-1}$ being the spectral radius of $\mathcal{A}_{1} \mathcal{D}_{0}$. Moreover, $\lambda_{0}$ is bounded from below by $\lambda_{0}^{*}$ and $\lambda_{1}^{*}$,
$\lambda_{0}^{*}=\left(2 \int_{0}^{\infty}\left(1-\prod_{k=1}^{K}\left(1-\exp \left(\alpha_{k} t\right)\right)^{C_{k}}\right) d t\right)^{-1} \geq\left(2 \sum_{k=1}^{K} \sum_{\ell=1}^{C_{k}} \frac{1}{\ell \alpha_{k}}\right)^{-1}=\lambda_{1}^{*}$.
Proof. Multiplying (3.9) by $\mathcal{D}_{0}$ and invoking (3.11) yields,

$$
\boldsymbol{\pi}^{(\lambda)}\left(\mathcal{A}_{0}+\lambda \mathcal{A}_{1}\right) \mathcal{D}_{0}=\boldsymbol{\pi}^{(\lambda)}\left(\boldsymbol{\Pi}_{0}-\mathcal{I}\right)+\boldsymbol{\pi}^{(\lambda)} \lambda \mathcal{A}_{1} \mathcal{D}_{0}=0
$$

Moreover, we have $\boldsymbol{\pi}^{(\lambda)} \boldsymbol{\Pi}_{0}=\boldsymbol{\pi}^{(\lambda)} \mathbf{e}^{\prime} \boldsymbol{\pi}^{(0)}=\boldsymbol{\pi}^{(0)}$, such that,

$$
\boldsymbol{\pi}^{(\lambda)}\left(\mathcal{I}-\lambda \mathcal{A}_{1} \mathcal{D}_{0}\right)=\pi^{(0)}
$$

The spectral radius of $\lambda \mathcal{A}_{1} \mathcal{D}_{0}$ is $\lambda / \lambda_{0}$. Hence for $\lambda<\lambda_{0},\left(\mathcal{I}-\lambda \mathcal{A}_{1} \mathcal{D}_{0}\right)$ is invertible and the Neumann series converges to the inverse,

$$
\sum_{k=0}^{\infty}\left(\lambda \mathcal{A}_{1} \mathcal{D}_{0}\right)^{k}=\left(\mathcal{I}-\lambda \mathcal{A}_{1} \mathcal{D}_{0}\right)^{-1}
$$

Combining the previous expressions immediately yields the series expansion (3.12).
As all elements but the first column of $\Pi_{0}$ are zero, only the first column of $\mathcal{D}_{0}$ may contain negative values; see $\sqrt{3.10}$. Moreover, the row sums of $\mathcal{D}_{0}$ are zero, hence the first column is equal in absolute value to the sum of the other columns. The entries in the first column of $\mathcal{D}_{0}$ have the following interpretation,

$$
\left[\mathcal{D}_{0}\right]_{\mathbf{x o}}=-\int_{0}^{\infty}\left(1-\left[\mathcal{P}_{0}(t)\right]_{\mathbf{x o}}\right) d t=-\mathrm{E}\left[T_{\mathbf{x}}\right]
$$

where $T_{\mathbf{x}}$ is a random variable denoting the time it takes to reach the empty state $\mathbf{0}$ from state $\mathbf{x}$ (assuming no arrivals). This interpretation shows that $\gamma \doteq \mathrm{E}\left[T_{\mathbf{c}}\right] \geq$ $\mathrm{E}\left[T_{\mathbf{x}}\right]$ for all $\mathbf{x} \in \mathcal{X}$ where $\mathbf{c}$ denotes the full state.

We have the following upper bound for $\gamma$. It is easy to see that $\gamma$ decreases if $\mu$ increases. Therefore, consider the system without service, that is with $\mu=0$. Then each customer in the $k$ th queue leaves at a rate $\alpha_{k}$ and the bound for $T_{\mathbf{c}}$ is the maximum of $\sum_{k} C_{k}$ independent exponentially distributed random variables. Hence, the corresponding cumulative distribution is the product of exponential distributions. The bound $\gamma_{0}^{*}$ for $\gamma$ is calculated by integrating this distribution,

$$
\gamma \leq \gamma_{0}^{*}=\int_{0}^{\infty}\left(1-\prod_{k=1}^{K}\left(1-e^{-\alpha_{k} t}\right)^{C_{k}}\right) d t
$$

Moreover, as the maximum of $K$ non-negative random variables is bounded from above by the sum of these random variables, we have the following crude upper bound for $\gamma_{0}^{*}$ (and $\gamma$ ),

$$
\begin{equation*}
\gamma \leq \gamma_{0}^{*} \leq \gamma_{1}^{*}=\sum_{k=1}^{K} \sum_{\ell=1}^{C_{k}} \frac{1}{\ell \alpha_{k}} \tag{3.13}
\end{equation*}
$$

the $k$ th term in the sum on the right-hand side corresponding to the mean time to deplete the $k$ th queue.

As the row sums of $\mathcal{A}_{1}$ are zero $\left(\mathcal{A}^{(\lambda)}\right.$ is a generator matrix for every $\lambda$ ), we have $\mathcal{A}_{1} \Pi_{0}=0$. Moreover, for any induced matrix norm, we have $\left\|\mathcal{A}_{1} \mathcal{D}_{0}\right\| \geq \lambda_{0}$. Therefore, we find,

$$
\lambda_{0}^{-1} \leq\left\|\mathcal{A}_{1} \mathcal{D}_{0}\right\|=\left\|\mathcal{A}_{1}\left(\mathcal{D}_{0}+\gamma \boldsymbol{\Pi}_{0}\right)\right\| \leq\left\|\mathcal{A}_{1}\right\|\left\|\mathcal{D}_{0}+\gamma \boldsymbol{\Pi}_{0}\right\| .
$$

In particular, using the maximum absolute row sum norm, we have $\left\|\mathcal{A}_{1}\right\|=$ $\sum_{k=1}^{K} \kappa_{k} \doteq \kappa ;\left[\mathcal{A}_{1}\right]_{\mathbf{x x}}=-\kappa$ if all queues are non-full in state x and $\left[\mathcal{A}_{1}\right]_{\mathbf{x x}}>-\kappa$
if this not the case. In view of the definition of $\gamma$, one easily verifies that the matrix $\mathcal{D}_{0}+\gamma \boldsymbol{\Pi}_{0}$ has no negative entries. Recalling that $\mathcal{D}_{0}$ has zero row sums, this shows that all row sums of $\mathcal{D}_{0}+\gamma \boldsymbol{\Pi}_{0}$ equal $\gamma:\left\|\mathcal{D}_{0}+\gamma \boldsymbol{\Pi}_{0}\right\|=\gamma$ and,

$$
\frac{1}{\lambda_{0}} \leq 2 \kappa \gamma \leq 2 \kappa \gamma_{0}^{*} \doteq \frac{1}{\lambda_{0}^{*}}
$$

which proves the lower bound $\lambda_{0}^{*}$ for $\lambda_{0}$. The lower bound $\lambda_{1}^{*}$ follows from $\lambda_{0}^{-1} \leq 2 \kappa \gamma$ and the crude bound (3.13) for $\gamma$.

Remark 5. The former theorem establishes a lower bound for the region of convergence of the series expansion. The existence of the series expansion in an interval around $\lambda=0$ can be established more easily. Indeed, by Cramer's rule, one directly verifies that the stationary probabilities are rational functions of $\lambda$. The region of convergence of the Maclaurin series expansion is therefore determined by the zero of the denominator with the smallest absolute value, which is distinct from 0 . As for every positive real $\lambda$ the stationary probability is between 0 and 1 , one further notes that this smallest zero is definitely not real and positive.

### 3.3 Overload analysis

We now study the system in overload. To this end, let $\alpha_{i}=\beta_{i} \nu$. The system of equations (3.1) then has the following matrix representation,

$$
\begin{equation*}
\pi \mathcal{A}=\pi\left(\widehat{\mathcal{A}}_{0}+\mu \widehat{\mathcal{A}}_{1}+\nu \widehat{\mathcal{A}}_{2}\right)=0 \tag{3.14}
\end{equation*}
$$

where the matrices $\widehat{\mathcal{A}}_{0}, \widehat{\mathcal{A}}_{1}$ and $\widehat{\mathcal{A}}_{2}$ neither depend on $\mu$ nor on $\nu$, and where we assume the states in the stationary vector $\boldsymbol{\pi}$ are ordered lexicographically. The matrix $\widehat{\mathcal{A}}_{0}$ contains transition rates corresponding to arrivals, and is an upper triangular matrix. Further, $\mathcal{A}_{1}$ only contains transition rates corresponding to departures, while $\mathcal{A}_{2}$ only contains transitions corresponding to abandonment.

We introduce the bivariate series expansion of the stationary probabilities $\pi(\mathbf{x})$ and of the corresponding stationary vector $\boldsymbol{\pi}$,

$$
\pi(\mathbf{x})=\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \pi_{m, n}(\mathbf{x}) \mu^{m} \nu^{n}, \quad \boldsymbol{\pi}=\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \boldsymbol{\pi}_{m, n} \mu^{m} \nu^{n}
$$

Plugging the expansion of the stationary vector above in 3.14) and isolating terms in $\mu^{m} \nu^{n}$ then yields,

$$
\begin{equation*}
\pi_{m, n} \widehat{\mathcal{A}}_{0}=-\pi_{m-1, n} \widehat{\mathcal{A}}_{1}-\pi_{m, n-1} \widehat{\mathcal{A}}_{2} \tag{3.15}
\end{equation*}
$$

and,

$$
\begin{align*}
\pi_{m, 0} \widehat{\mathcal{A}}_{0} & =-\pi_{m-1,0} \widehat{\mathcal{A}}_{1} \\
\pi_{0, n} \widehat{\mathcal{A}}_{0} & =-\pi_{m, n-1} \widehat{\mathcal{A}}_{2} \\
\pi_{0,0} \widehat{\mathcal{A}}_{0} & =0 \tag{3.16}
\end{align*}
$$

for $m, n \in \mathbb{N}^{+}$. Moreover, by the normalisation condition $\boldsymbol{\pi} \mathbf{e}^{\prime}=1$, we find,

$$
\pi_{0,0} \mathbf{e}^{\prime}=1, \quad \pi_{m, n} \mathbf{e}^{\prime}=0
$$

for $(m, n) \in \mathbb{N}^{2} \backslash\{(0,0)\}$. Recalling the triangularity of $\widehat{\mathcal{A}}_{0}$, the recursive equations 3.15 - 3.16 can be readily solved. For convenience, we express the recursion in terms of the system parameters below.

### 3.3.1 Numerical series expansion

First, as for the light-traffic case, $\boldsymbol{\pi}_{0,0}$ is trivial as all queues eventually become full when there are no departures,

$$
\pi_{0}(\mathbf{x})= \begin{cases}1 & \text { for } x_{1}=C_{1}, \ldots, x_{K}=C_{K} \\ 0 & \text { otherwise }\end{cases}
$$

We can again calculate the higher order terms recursively. Given the values for $m+n<k$, we find the terms for $m+n=k$ by evaluating the equations below in lexicographical order. For $\mathbf{x} \in \mathcal{X} \backslash\{\mathbf{c}\}$ with $\mathbf{c}=\left[C_{1}, C_{2}, \ldots, C_{K}\right]$, we have,

$$
\begin{align*}
& \pi_{m, n}(\mathbf{x})=\left(\sum_{k=1}^{K} \lambda_{k} 1_{\left\{x_{k}<C_{k}\right\}}\right)^{-1} \\
& \times\left(-\pi_{m, n-1}(\mathbf{x}) \sum_{k=1}^{K} \beta_{k}\left(x_{k}-E(\mathbf{x})\right)-\pi_{m-1, n}(\mathbf{x}) E(\mathbf{x})+\pi_{m-1, n}(\mathbf{x}+\mathbf{e})\right. \\
& \left.+\sum_{k=1}^{K} \pi_{m, n}\left(\mathbf{x}-\mathbf{e}_{k}\right) \lambda_{k}+\sum_{k=1}^{K} \pi_{m, n-1}\left(\mathbf{x}+\mathbf{e}_{k}\right) \beta_{k}\left(x_{k}+1-E\left(\mathbf{x}+\mathbf{e}_{k}\right)\right)\right) \tag{3.17}
\end{align*}
$$

whereas for $\mathbf{x}=\mathbf{c}$ we have,

$$
\pi_{m, n}(\mathbf{c})=-\sum_{\mathbf{x} \in \mathcal{X} \backslash\{\mathbf{c}\}} \pi_{m, n}(\mathbf{x})
$$

Remark 6. In contrast to the light-traffic approach, the numerical complexity is now $O\left(C^{K} K N^{2}\right)$, as the calculation of every value in $\pi_{n, m}(\mathbf{x})$ is $O(K)$. The algorithm is therefore considerably slower than the light-traffic approximation for
large $N$. In addition, more memory is required as well. In the light-traffic approximation it is sufficient to keep track of the last term in the expansion only. Now, calculating the $(m, n)$ terms with $m+n=k$ requires all $(m, n)$ terms in the expansion with $m+n=k-1$.

As for the light-traffic expansion, the number of non-zero terms in the vector $\boldsymbol{\pi}_{m, n}$, is considerably smaller than the length of the vector. By the $n$-event rule, $\boldsymbol{\pi}_{m, n}(\mathbf{x})$ is only non-zero for states $\mathbf{x}$ that can be reached from state $\mathbf{c}$ by at most $m$ departures by impatience and $n$ departures upon service completion. Accounting for this observation, the numerical complexity reduces to $O\left(\min (C, K)^{K} K N^{2}\right)$.

Likewise, if the abandonment and arrival rates are the same for all queues, one can again exploit the symmetry: $\pi_{m, n}(\mathbf{x})=\pi_{m, n}(\mathbf{y})$ for any permutation $\mathbf{y}$ of $\mathbf{x}$. Remark 7. The approach for light traffic can be adopted to study the system in overload as well. To this end, one scales the abandonment rates with $\mu, \alpha_{i}=\beta_{i} \mu$ and investigates the series expansion in $\mu=0$. Scaling the abandonment rates with $\mu$ implies that there are no (lexicographically) downward transitions for $\mu=0$. In other words, the generator matrix of the Markov chain for $\mu=0$ is triangular and the light-traffic approach applies.

### 3.3.2 Closed form expressions for the symmetric coupled queueing system

For the symmetric coupled queueing system we obtain closed-form expressions for the 2 nd order expansion of the first two moments of the queue content. As the system is symmetric we have $\alpha_{k}=\alpha$ and $\lambda_{k}=\lambda$ for $k=1, \ldots, K$. In addition, we assume that the queue capacities are equal $C_{k}=C$ for all $k=1, \ldots, K$ and that $C>K>2$. By repeated application of 3.17, we then have the following second order approximation for the first two moments of the queue content $X$ :

$$
\begin{aligned}
\mathrm{E}[X] & \approx C-\frac{C-1}{\lambda} \alpha-\frac{1}{\lambda} \mu-\frac{(C-1)(C-3)}{\lambda^{2}} \alpha^{2}-2 \frac{C-2}{\lambda^{2}} \alpha \mu-\frac{1}{\lambda^{2}} \mu^{2}, \\
\mathrm{E}\left[X^{2}\right] & \approx C^{2}-\frac{(2 C-1)(C-1)}{\lambda} \alpha-\frac{2 C-1}{\lambda} \mu-\frac{(2 C-7)(C-1)^{2}}{\lambda^{2}} \alpha^{2} \\
& -\frac{(2 C-5)(C-1)}{\lambda^{2}} \alpha \mu-\frac{2 C-3}{\lambda^{2}} \mu^{2} .
\end{aligned}
$$

### 3.4 Fluid limit

In this section, we develop a fluid limit for the queueing model at hand. We hereby make the following additional assumptions: the abandonment rates $\alpha_{k}$ are nonzero, the arrival rates $\lambda_{k}$ in all queues exceed the service rate, i.e. $\lambda_{k}>\mu$, and all queue capacities $C_{k}$ are infinite. We consider the scaling:

$$
\alpha_{k} \mapsto \alpha_{k}, \quad \lambda_{k} \mapsto N \lambda_{k}, \quad \mu \mapsto N \mu .
$$

The infinite capacity assumption is relaxed below. We will indicate how to adapt the proof to the case of finite capacities, provided that they are scaled as follows $C_{k} \mapsto N C_{k}$, and satisfy:

$$
\begin{equation*}
C_{k}>\frac{\lambda_{k}-\mu}{\alpha_{k}} \tag{3.18}
\end{equation*}
$$

for $k=1,2, \ldots, K$.
Recalling that $X_{k}(t)$ denotes the number of customers in the $k$ th queue at time $t$, let $X_{k}^{N}(t)$ be the number of customers in the $k$ th queue at time $t$ for the system with arrival rates $N \lambda_{k}$ and service rate $N \mu$. In the spirit of the monograph of Ethier and Kurtz [30], we express the evolution of the system in terms of Poisson processes with random time changes:

$$
\begin{aligned}
& X_{k}^{N}(t)=X_{k}^{N}(0)+Y_{k}\left(N \lambda_{k} t\right) \\
&-Z_{k}\left(\alpha_{k} \int_{0}^{t} X_{k}^{N}(s) d s\right)-U\left(N \mu \int_{0}^{t} \prod_{k} 1_{\left\{X_{k}^{N}(s)>0\right\}} d s\right),
\end{aligned}
$$

where $Y_{k}(\cdot), Z_{k}(\cdot)$ and $U(\cdot)$ are independent Poisson processes with unit rate. We further assume that the random variables $X_{k}^{N}(0) N^{-1}$ converge to the deterministic constants $\rho_{k}(0)>0$ for $N \rightarrow \infty$.

We will show that the process has the following fluid limit:

$$
\begin{equation*}
\rho_{k}(t)=\frac{\lambda_{k}-\mu}{\alpha_{k}}\left(1-e^{-\alpha_{k} t}\right)+\rho_{k}(0) e^{-\alpha_{k} t}, \tag{3.19}
\end{equation*}
$$

where we note that these functions can also be written as the unique solutions of the following integral equations:

$$
\rho_{k}(t)=\rho_{k}(0)+\left(\lambda_{k}-\mu\right) t-\alpha_{k} \int_{0}^{t} \rho_{k}(s) d s
$$

In order to establish the fluid limit, we want to prove that the processes $\hat{X}_{k}^{N}(t) \doteq$ $\left(N^{-1} X_{k}^{N}(t)-\rho_{k}(t)\right)$, converge to zero processes, that is, that

$$
\sup _{t \in[0, T]} \sum_{k}\left|\hat{X}_{k}^{N}(t)\right|
$$

converges to 0 in probability as $N \rightarrow \infty$. We prove this proposition by making use of Grönwall's lemma and of the functional law of large numbers for Poisson processes. Let us rewrite the expression for $\hat{X}_{i}^{N}(t)$ as follows:
$\hat{X}_{k}^{N}(t)=\hat{X}_{k}^{N}(0)+M_{1, k}^{N}(t)-M_{2}^{N}(t)-M_{3}^{N}(t)-M_{4, k}^{N}(t)-\alpha_{k} \int_{0}^{t} \hat{X}_{k}^{N}(s) d s$,
with,

$$
\begin{aligned}
& M_{1, k}^{N}(t)=N^{-1} Y_{k}\left(N \lambda_{k} t\right)-\lambda_{k} t \\
& M_{2, k}^{N}(t)=N^{-1} Z_{k}\left(\alpha_{k} \int_{0}^{t} X_{k}^{N}(s) d s\right)-N^{-1} \alpha_{k} \int_{0}^{t} X_{k}^{N}(s) d s \\
& M_{3}^{N}(t)=N^{-1} U\left(N \mu \int_{0}^{t} \prod_{k} 1_{\left\{X_{k}^{N}(s)>0\right\}} d s\right)-\mu \int_{0}^{t} \prod_{k} 1_{\left\{X_{k}^{N}(s)>0\right\}} d s, \\
& M_{4}^{N}(t)=\mu \int_{0}^{t} \prod_{k} 1_{\left\{X_{k}^{N}(s)>0\right\}} d s-\mu t .
\end{aligned}
$$

We immediately see that

$$
\begin{aligned}
\sum_{k}\left|\hat{X}_{k}^{N}(t)\right| \leq & \sum_{k}\left|\hat{X}_{k}^{N}(0)\right|+\sum_{k} \sup _{t \in[0, T]}\left|M_{1, k}^{N}(t)\right|+\sum_{k} \sup _{t \in[0, T]}\left|M_{2, k}^{N}(t)\right| \\
& +\sup _{t \in[0, T]}\left|M_{3}^{N}(t)\right|+\sup _{t \in[0, T]}\left|M_{4}^{N}(t)\right|+\alpha^{*} \int_{0}^{t} \sum_{k}\left|\hat{X}_{k}^{N}(s)\right| d s
\end{aligned}
$$

where $\alpha^{*}$ is the largest $\alpha_{i}$. Using the integral form of Grönwall's lemma, we get

$$
\begin{array}{r}
\sum_{k}\left|\hat{X}_{k}^{N}(t)\right| \leq\left(\sum_{k}\left|\hat{X}_{k}^{N}(0)\right|+\sum_{k} \sup _{t \in[0, T]}\left|M_{1, k}^{N}(t)\right|+\sum_{k} \sup _{t \in[0, T]}\left|M_{2, k}^{N}(t)\right|\right. \\
\left.+\sup _{t \in[0, T]}\left|M_{3}^{N}(t)\right|+\sup _{t \in[0, T]}\left|M_{4}^{N}(t)\right|\right) \exp \left(\alpha^{*} t\right) .
\end{array}
$$

Hence, to establish the fluid limit, it suffices to show that the five terms between parentheses converge to zero in probability.

The convergence of $\left|\hat{X}_{k}^{N}(0)\right|$ is by assumption, while the convergence of $\left|M_{1, k}^{N}(t)\right|$ is a standard application of the functional law of large numbers for Poisson processes.

Regarding $\left|M_{2, k}^{N}(t)\right|$, observe that from the inequality $X_{k}^{N}(t) \leq X_{k}^{N}(0)+$ $Y_{k}\left(N \lambda_{k} t\right)$ we have
$\lim _{N \rightarrow \infty} \frac{1}{N} \int_{0}^{t} X_{k}^{N}(s) d s \leq \lim _{N \rightarrow \infty} \frac{1}{N} \int_{0}^{t} X_{k}^{N}(0)+Y_{k}\left(N \lambda_{k} s\right) d s=\rho_{k}(0) t+\frac{1}{2} \lambda_{k} t^{2}$.
Hence, for any fixed $\epsilon>0$, we have the crude inequality

$$
N^{-1} \int_{0}^{T} X_{k}^{N}(s) d s \leq\left(\rho_{k}(0)+\frac{1}{2} \lambda_{k} T\right) T+\epsilon \doteq \hat{T}
$$

on a set of at least probability $1-\epsilon$ for $N$ large enough. We apply the functional limit of large numbers for Poisson processes on the processes $Z_{k}$ in the interval
$[0, \hat{T}]$, and by force we also have convergence of the original term. The same reasoning applies to the term $\left|M_{3}^{N}(t)\right|$. In this case, we use the deterministic upper bound

$$
\mu \int_{0}^{t} \prod_{k} 1_{\left\{X_{k}^{N}(s)>0\right\}} d s \leq \mu t
$$

For the convergence of $\left|M_{4}^{N}(t)\right|$ to hold, we must establish that for large enough $N$, the queues stay non-empty in the entire interval $[0, T]$. To do so, note that we have $X_{k}^{N}(t) \geq \tilde{X}_{k}^{N}(t)$, where the process $\tilde{X}_{k}^{N}(t)$ is defined as

$$
\tilde{X}_{k}^{N}(t)=X_{k}^{N}(0)+Y_{k}\left(N \lambda_{k} t\right)-Z_{k}\left(\alpha_{k} \int_{0}^{t} X_{k}^{N}(s) d s\right)-U(\mu t) .
$$

Using the same arguments as above, we can show that this process converges to the same fluid limit $\left\{\rho_{k}(t)\right\}$, the bound on the last term in $\tilde{X}_{k}^{N}(t)$ now being immediate by the functional strong law of large numbers for Poisson processes. As $\rho_{k}(t)>0$ if $\rho_{k}(0)>0$, it then follows that the larger process $X_{k}^{N}(t)$ must also stay away from zero.
Remark 8. The reasoning for the last term can be repeated to establish the fluid limit for the kitting process with finite capacity buffers. Indeed, the process with infinite queue capacity is an upper bound for any system where one or more of the queue capacities is finite. One then only needs to show that the fluid limit stays away from the boundaries $C_{k}$. This is indeed the case by equation 3.18.

### 3.5 Numerical results and discussion

Having established 3 approximations for the coupled queueing system, we now investigate the accuracy of the proposed approximations by some numerical examples.

We first focus on the coupled queueing system in light traffic. Figures 3.2 and 3.3 show the light traffic approximations of the mean and variance of the queue content for a symmetric kitting process with 5 queues, each having capacity $C=$ 10. The service rate is $\mu=1$ and the abandonment rate is the same in all queues - $\alpha_{i}=\alpha$ for $i=1, \ldots, 5$ - with $\alpha=0.1$ or $\alpha=0.2$ as indicated. We compare the 3rd, 10th and 50th order approximations, and additionally also simulate the system for verifying the accuracy of the approximations.

For the symmetric system at hand, the 3rd order approximation equals the first order approximation for the mean and the second order approximation for the variance; the explicit expressions (3.7) and (3.8) show that the coefficients of order 2,3 and 4 for the mean and of order 3 and 4 for the variance are equal to zero (as we have 5 queues). The 10th order approximation of the mean queue content is already quite good, while this is not the case for the 10th order approximation of


Figure 3.2: Order 3, 10 and 50 light traffic approximation for the mean queue content of the kitting system with service rate $\mu=1$, and with $K=5$ queues, each having capacity $C=10$ and abandonment rates $\alpha=0.1$ or $\alpha=0.2$ as indicated.


Figure 3.3: Order 3, 10 and 50 light traffic approximation for the variance of the queue content of the kitting system with service rate $\mu=1$, and with $K=5$ queues, each having capacity $C=10$ and abandonment rates $\alpha=0.1$ or $\alpha=0.2$ as indicated.
the variance for abandonment rate $\alpha=0.1$. Higher order approximations improve on the accuracy for sufficiently small $\lambda$. For high $N$, we obtain very accurate results, up to a certain $\lambda_{\max }$ where the series expansion no longer converges to the correct result. Moreover, we have the same $\lambda_{\max }$ for the mean and variance approximations. The sudden deviation of the correct value is an indicator that this $\lambda_{\text {max }}$ corresponds to the radius of convergence of the series expansion.

Assuming the same parameters as in Figures 3.2 and 3.3 . Figures 3.4 and 3.5 depict the mean and variance of the queue content for the symmetric kitting process for higher values of $\alpha$ : for $\alpha=1$ and $\alpha=2$. We again compare the 3 rd, 10th and 50th order approximations, and simulate the system for verifying the accuracy of the approximations. For these high values of $\alpha$, it is hard to discern the mean


Figure 3.4: Order 3, 10 and 50 light traffic approximation for the mean queue content of the kitting system with service rate $\mu=1$, and with $K=5$ queues, each having capacity $C=10$ and abandonment rates $\alpha=1$ or $\alpha=2$ as indicated.


Figure 3.5: Order 3, 10 and 50 light traffic approximation for the variance of the queue content of the kitting system with service rate $\mu=1$, and with $K=5$ queues, each having capacity $C=10$ and abandonment rates $\alpha=1$ or $\alpha=2$ as indicated.
value and the variance plots. This can be explained as follows. For high $\alpha$, the abandonment process dominates the service process and the kitting process can be approximated by a system of parallel $M / M / \infty$ queues (the abandonment process being the service process of the $M / M / \infty$ queues). It is well known that the queue content distribution of an $M / M / \infty$ process is a Poisson distribution, the Poisson distribution having equal mean and variance.

The accuracy of the overload approximations is illustrated by Figures 3.6 and 3.7 that depict the mean and variance of the queue content, vs. the service rate $\mu$. As for light traffic, we show the 3rd, 10th and 50 th order approximations and include simulation results to assess the accuracy of the approximations. We again consider a system with 5 queues. The arrival rate $\lambda_{k}$ is 1 for all queues, whereas the


Figure 3.6: Order 3, 10 and 50 heavy-traffic approximation for the mean queue content of the kitting system with arrival rate $\lambda=1$, and with $K=5$ queues, each having capacity $C=10$ and abandonment rates $\alpha=0, \alpha=0.05$ or $\alpha=0.1$ as indicated.


Figure 3.7: Order 3, 10 and 50 heavy traffic approximation for the variance of the queue content of the kitting system with arrival rate $\lambda=1$, and with $K=5$ queues, each having capacity $C=10$ and abandonment rates $\alpha=0, \alpha=0.05$ or $\alpha=0.1$ as indicated.
abandonment rate is $\alpha$ for every queue, different values of $\alpha$ being considered as depicted. As for the light traffic approximation, we find a reasonable accuracy of lower order approximations and accurate results for higher-order approximations for $\mu$ up to a specific value $\mu_{\max }$, while the series expansion no longer converges to the correct result for larger $\mu$. This again is an indicator that $\mu_{\max }$ corresponds to the radius of convergence of the series expansion.

Note that the overload approximation is a bivariate expansion. While the approximation for $\lambda=0$ is exact for the light traffic approximation, this is not the case for $\mu=0$ in the overload expansion. Indeed, the approximation is only exact for $\mu=\alpha=0$ and we evaluate for non-zero $\alpha$. This is readily observed for the $3 r d$ order approximations of mean and variance in Figures 3.6 and 3.7. respectively.


Figure 3.8: Mean queue content versus the abandonment rate for a kitting process with $K=5$ queues with $\lambda=1$. The queue capacity is $C=40, C=60$ and $C=80$ as indicated whereas the service rate is $\mu=0.1$ or $\mu=0.25$ as indicated.

Finally Figure 3.8 depicts the mean queue content versus the abandonment rate $\alpha$. There are 5 queues, the arrival rate is $\lambda=1$ for all queues, and the abandonment rate $\alpha$ and queue capacity $C$ are equal for all queues. We consider different sizes of the queue capacity $C$ and service rates $\mu=0.1$ and $\mu=0.25$. As the system is in overload, both the overload approximation and the fluid approximation can be used. Figure 3.8 depicts both approximations, as well as simulation results to verify the accuracy of the approximations. For large $\alpha$, one observes that the fluid approximation is accurate while this is not the case for small $\alpha$. Indeed, the constraint on the queue capacity (3.18) for the fluid approximation implies that $\alpha$ should be at least $(\lambda-\mu) / C$. In contrast to the fluid approximation, the overload approximation is most accurate for small $\alpha$. As illustrated by Figure 3.8 , both approximations are complementary. Indeed, the simulation results reveal that the combined approximation is accurate for all $\alpha$.

### 3.6 Conclusions

We considered a numerical technique based on Maclaurin series expansions to study a coupled queueing system with customer impatience. For the light-traffic approximation, we noted that the series expansion technique resembles the GaussSeidel method, while it delivers an approximation in a range of the parameter space. The overload approximation introduces a bivariate series expansion, expressing the performance measures of interest as a bivariate polynomial of the service rate and the scaling factor of the abandonment rates. While the bivariate series expansion is computationally more expensive, we found accurate approximations in reasonable time. Although the prime aim of the series approximations was the
development of a fast approximation algorithm, we also included expressions for the $K$ th order light traffic approximation for the symmetric coupled queueing system with $K$ queues, as well as the 2 nd order approximation for the symmetric system in overload. Finally, we also studied and formally proved the fluid limit of the coupled queueing system when the system is in overload. Numerical experiments particularly revealed that a combination of the overload approximation and the fluid limit allows for approximating the system in the complete range of the abandonment rate $\alpha$ when the arrival rate exceeds the service rate.

## References

[1] D. Beyer, F. Cheng, S.P. Sethi, M. Taksar. Markovian Demand Inventory Models, Springer, 2010.
[2] G. Liberopoulos, C.T. Papadopoulos, B. Tan, J.M. Smith, S.B. Gershwin (Eds.). Stochastic Modeling of Manufacturing Systems. Springer, 2006.
[3] F. Ju, J. Li, J.A. Horst. Transient analysis of serial production lines with perishable products: Bernoulli reliability model. IEEE Transactions on automatic control, 62(2):694-707, 2017.
[4] J. Harrison. Assembly-like queues. Journal Of Applied Probability, 10:354367, 1973.
[5] W.J. Hopp, J.T. Simon. Bounds and heuristics for assembly-like queues. Queueing Systems, 4:137-156, 1989.
[6] G. Latouche. Queues with paired customers. Journal of Applied Probability, 18(3)684-696, 1981.
[7] R. Ramakrishnan, A. Krishnamurthy. Analytical approximations for kitting systems with multiple inputs. Asia-Pacific Journal of Operations Research, 25(2):187-216, 2008.
[8] R. Ramakrishnan, A. Krishnamurthy. Performance evaluation of a synchronization station with multiple inputs and population constraints. Computers \& Operations Research, 39:560-570, 2012.
[9] R. Nelson, A.N. Tantawi. Approximate analysis of fork join synchronization in parallel queues. IEEE Transactions on Computers, 37(6):739-743, 1988.
[10] Z. Qiu, Zhan, J.F. Perez, P.G. Harrison. Beyond the mean in fork-join queues: Efficient approximation for response-time tails. Performance Evaluation, 91:99-116, 2015.
[11] S. Borst, O. Boxma, M. van Uitert. The asymptotic workload behavior of two coupled queues. Queueing Systems, 43(1-2):81-102, 2003.
[12] C. Knessl, J.A. Morrison. Asymptotic Analysis of Two Coupled Queues with Vastly Different Arrival Rates and Finite Customer Capacities. Studies in Applied Mathematics, 128(2):107-143, 2012.
[13] J. Pender. An analysis of nonstationary coupled queues. Telecommunication Systems, 61(4):823-838, 2016.
[14] J. Resing, L. Ormeci. A tandem queueing model with coupled processors. Operations Research Letters, 31(5):383-389, 2003.
[15] S. Borst, M. Jonckheere, L. Leskela. Stability of parallel queueing systems with coupled service rates. Discrete Event Dynamic Systems-Theory and Applications, 18(4):447-472, 2008.
[16] M. Takahashi, H. Osawa, T. Fujisawa. On a synchronization queue with two finite buffers. Queueing Systems, 36:107-23, 2000.
[17] E. De Cuypere, D. Fiems. Performance evaluation of a kitting process. In: Proceedings of the 18th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2011), pages 175-188, Venice, June 2011.
[18] E. De Cuypere, K. De Turck, D. Fiems. A Maclaurin-series expansion approach to multiple paired queues. Operations Research Letters, 42(3):203207, 2014.
[19] E. Altman, K.E. Avrachenkov, R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. Advances in Applied Probability, 36(3):839-853, 2004.
[20] J.B. Lasserre. A formula for singular perturbations of Markov chains. Journal of Applied Probability, 31(3):829-833, 1994.
[21] K.E. Avrachenkov, J.A. Filar, P.G. Howlett. Analytic perturbation theory and its applications. SIAM, 2013.
[22] W.B. van den Hout. The power-series algorithm: a numerical approach to Markov processes. PhD Thesis. Tilburg University, 1996.
[23] G. Koole. On the power series algorithm. CWI, 1994.
[24] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In Performance Evaluation of Computer and Communication Systems, pages 53-80, 1993.
[25] J.P.C. Blanc, R.D. van der Mei. Optimization of polling systems with Bernoulli schedules. Performance Evaluation, 22(2):139-158, 1995.
[26] B. Błaszczyszyn, T. Rolski, V. Schmidt. Advances in Queueing: Theory, Methods and Open Problems, chapter Light-traffic approximations in queues and related stochastic models. CRC Press, Boca Raton, Florida, 1995.
[27] I. Kovalenko. Rare events in queueing theory. A survey. Queueing systems, 16(1):1-49, 1994.
[28] P.J. Schweitzer. Perturbation Theory and Finite Markov Chains, Journal of Applied Probability, 5(2):401-413, 1968.
[29] The Online Encyclopedia of Integer Sequences. https://oeis.org/ A000070
[30] S.N. Ethier, T.G. Kurtz. Markov Processes: Characterization and Convergence. Wiley and Sons, Second Edition. 2005.
[31] R.W.R. Darling, J.R. Norris. Differential equation approximations for Markov chains. Probability Surveys, 5:37-79, 2008.

## 4

# Queueing Analysis of Opportunistic Scheduling with Spatially Correlated Channels 

Ekaterina Evdokimova, Koen De Turck, Sabine Wittevrongel and Dieter Fiems


#### Abstract

Exploiting differences in supported transmission rates between mobile users, opportunistic scheduling promises a substantial increase of the aggregate throughput of wireless networks. In this paper, we present a Markov model to study the trade-off between fairness and wireless efficiency of opportunistic scheduling at an access point serving multiple mobile users. The Markovian model description includes both the state of the queues and the state of the wireless channel of the different mobile users. The size of the state space of the Markov model at hand preventing a direct solution, we develop a numerical analysis technique based on Maclaurin series expansions to solve the system in light traffic and in overload. We illustrate the accuracy of our approach by numerical examples.


### 4.1 Introduction

Efficiently allocating networking resources is key for the performance of many multi-user (MU) wireless communication systems. Such allocations aim at optimizing performance metrics like network throughput, delay and jitter, while at the same time retaining fairness between the users [1]. In contrast to wireline networks, the transmission rates supported by individual mobile users vary over time and per user. For example, the channel capacity of an individual user depends on its distance to the base station. In addition, users undergo short-term location-dependent fading, the statistics of such short-term variations differing across users [2]. Therefore, scheduling to the users with the best rates at any given time may allow for substantially increasing the aggregate throughput of the wireless network.

Opportunistic scheduling is a promising cross-layer method that holds the potential of significantly improving wireless networks' efficiency in the near future. The technique however immediately brings into focus the trade-off between wireless efficiency (i.e., a preference to schedule to the best channel) and fair scheduling (i.e., each user is entitled to a certain amount of network resources) [3]. Since the introduction of opportunistic scheduling in [4], numerous schedulers have been proposed for different instances of wireless networks, such as mobile cellular networks, cognitive radio, WiMAX, MIMO systems; see [5-12] and the references therein.

While holding the promise to increase the throughput, opportunistic scheduling also faces some limitations. In order to select where to send to, the scheduler requires accurate channel state information (CSI). Such CSI is reported at the low rate of the feedback channel, and may therefore be outdated at the time the scheduler decides where to send to. In other words, the opportunistic scheduler always has to rely on partially known channel state information, the timeliness depending on the feedback delay. The discussion how much feedback is sufficient in order to benefit from the MU diversity is actively investigated [13, 14]. Furthermore, the gain of opportunistic scheduling may suffer from cross-channel correlations between users [15]. Typically the wireless networks exploiting MU diversity achieve the maximum performance in case of independent channels while under dependent channels the overall transmission capacity drops [16, 17]. Finally, the scheduling policy itself cannot require excessive computational complexity. While literature often focuses on designing optimal schedulers for particular fairness and quality-of-service (QoS) requirements, such schedulers may require too much computations to be practically implementable. Therefore, heuristic approaches are often preferred in order to simplify the implementation [18, 19]. Although many schedulers have been proposed, for the next generation of wireless systems there is a demand for new policies that rely on less feedback information, account for the spa-
tial correlations between users and require low computational complexity, while providing nearly optimal scheduling decisions. See for example [5, 8, 19- 22] for recent advances in opportunistic scheduling policies.

Along with developments in scheduling strategies, there is also an increasing demand for performance evaluation tools (i.e., instruments for testing, comparing and designing schedulers) for opportunistic schedulers. The presentation of an analytic framework for studying opportunistic schedulers is the main contribution of this paper. Specifically, we propose a method for the fast performance evaluation of wireless networks equipped with one access point (AP) serving multiple mobile users under varying transmission conditions.

Only few authors assess the performance of opportunistic schedulers by analytic means, most assessments of schedulers relying on simulations, see e.g. [22-24]. This is not surprising as stochastic models of opportunistic schedulers involve multiple queues. This results in a Markov model with a multidimensional state space. Even for a limited number of buffers (or mobile nodes) and limited buffer capacities, the state space of the Markov chain is huge which makes direct solution techniques numerically unfeasible. In [25], the authors propose a decomposition method to avoid the state space explosion problem. The approach relies on representing the MU system with $K$ mobile users as a deterministic and stochastic Petri net (DSPN), decomposed into $K$ subnets. Since the subnets are analyzed separately, the MU system is represented with far fewer states than the original Markov model, thereby achieving a low computational complexity. This approach rules out most interactions between the mobile users which is essential for a complete performance study. Indeed, the interaction is key for the scheduler as each allocation decision impacts all mobile users. A similar decomposition approach is presented in [26] for cognitive radio spectrum allocation. Here, a queueing model is analysed by matrix-analytical methods. However, the study mainly focuses on the single-queue case with an extrapolation to multiple queues. Finally, [19] studies the formation of time-space batches of packets assigned for simultaneous transmission. The authors do not track the number of packets for the different destinations, thereby again avoiding the inherent multidimensionality.

Our approach neither relies on decomposition nor on an extrapolation of the single-queue case, which enables us to accurately study the interactions among the queues. We consider a continuous-time Markovian model with a separate queueing state variable for each mobile node and with Rayleigh fading channels. Our channel model not only accounts for temporal correlation, but also for spatial correlation between the channels. The effects of fading are introduced in the queueing model by a transmission environment variable. The transmission environment is an exogenous continuous-time Markov process with a finite number of states in accordance with [27]. The overall queueing model at hand is a continuous-time Markov process. This means that opportunistic scheduling is modelled as an as-
signment of transmission rates to queues, thereby (possibly) accounting for both the current number of packets in the queue as well as the channel conditions. The transmission rate being proportional to the chance of a service time completion in an infinitely small interval, such a model still allows for some uncertainty in having a successful transmission or in the estimation of the channel conditions, but largely makes abstraction of the details of scheduling the packets.

The size of the state space of the Markov process at hand makes a direct solution technique computationally infeasible. For example, considering a system with 10 mobile nodes, and a buffer capacity of 10 per mobile node yields a state space with size exceeding $10^{10}$. To cope with such state spaces, we rely on Maclaurin series expansions of the solution of the Markov process [28-31] to assess the performance both fast and accurately. Depending on the context in which they are introduced, series expansion techniques for Markov chains are referred to as perturbation techniques, the power series method or light-traffic analysis. While the naming is not absolute, perturbation methods are mainly motivated by the assessment of the sensitivity of the performance measures with respect to a system parameter. The case where the perturbation does not preserve the class structure of the non-perturbed chain - the so-called singular perturbations - has received much attention in literature [28, 32]. The power series method transforms a Markov chain of interest in a set of Markov chains parametrised by a possibly artificial parameter. When the parameter is zero, the chain is not only easily solved, but one can also obtain the series expansion in the parameter. When the parameter is one, one gets the original Markov chain such that the series expansion can be used to approximate the solution of the original Markov chain, provided the radius of convergence of the series expansion exceeds one [33]. Finally, light-traffic analysis often corresponds to the series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in stochastic systems, we further refer the reader to the surveys in [34] and [35] and the recent book [36]. The present study most closely relates to the numerical series expansion approach of [30] and [31]. In contrast to this work, the present unperturbed chain is not upper-diagonal, but block upper-diagonal. It is shown below, that calculating the terms in the series expansion - in overload as well as under light traffic - is much easier than solving the queueing model for any particular load. The present paper extends our preliminary findings presented in [37, 38].

The remainder of the paper is organised as follows. The next section introduces the modelling assumptions and settles the notational conventions. The proposed analysis technique is then outlined in Section 4.3 Section 4.4 discusses the Markovian channel model assuming Rayleigh fading. In order to validate the proposed performance evaluation method we validate the accuracy of our results by simulation in section 4.5. For the sake of demonstration, we implement several simple schedulers [4, 31, 39, 40] and apply the proposed methodology for systems with


Figure 4.1: Queueing model for the opportunistic scheduler
spatially independent and correlated channels. Finally, conclusions are drawn in Section 4.6

### 4.2 Queueing model

We consider a wireless AP opportunistically sending packets to multiple mobile nodes. The AP is modelled as a Markovian queueing model with $K$ finite-capacity queues that share a common transmission channel, as depicted in Figure 4.1. Each queue corresponds to the AP buffer of a particular mobile node. Let $N_{k}(t)$ be the number of packets in the buffer of the $k$ th mobile node at time $t$, let $C_{k}$ be the capacity of this buffer and let $\mathbf{N}(t)=\left[N_{1}(t), \ldots, N_{K}(t)\right]$ be the vector with elements $N_{k}(t)$. Arrivals at the different buffers are modelled by independent Poisson processes; $\lambda_{k}$ denotes the arrival rate at queue $k$. We further assume that the packet sizes in the $k$ th queue are exponentially distributed with rate $\theta_{k}$.

The mobile nodes experience different time-varying channel conditions. To model variations of the channel conditions in both space and time, we introduce an exogenous continuous-time Markov process $M(t)$ that modulates the states of the different wireless transmission channels. We refer to $M(t)$ as the background process. Let the finite set $\mathcal{M}$ be the state space of this Markovian background process, let $M$ denote the cardinality of $\mathcal{M}$, and let $\alpha_{i j}$ denote the transition rate from state $i$ to state $j, i \neq j, i, j \in \mathcal{M}$. For every background state $m \in \mathcal{M}$, let $\mathbf{g}_{m}=\left[g_{m 1}, \ldots, g_{m K}\right]$ be a vector whose $k$ th element quantifies the channel conditions as experienced by the $k$ th mobile node. Hence, the channel condition vector at time $t$ is $\mathbf{G}(t):=\mathbf{g}_{M(t)}$. Without loss of generality, we assume that $g_{m k} \in[0 \ldots 1]$, where $g_{m k}=1$ corresponds to the best expected channel quality that allows transmission at the highest rate and $g_{m k}=0$ represents the case of poor channel quality when transmission is not feasible.

Given the channel conditions $\mathbf{G}(t)$ and the number of packets in the queues $\mathbf{N}(t)$, the opportunistic scheduler assigns service rates to the different queues. Let $\Psi_{k}(\mathbf{g}, \mathbf{n})$ be the rate assigned to the $k$ th mobile node, assuming channel conditions $\mathbf{G}(t)=\mathbf{g}$ and queueing state $\mathbf{N}(t)=\mathbf{n}$. We do not make any additional assump-
tions on the scheduling rule. Various specific schedulers are studied in section 4.5 including MaxRate, MaxWeight, Longest Connected Queue, and the Generalized and Differentiated processor sharing schedulers.

In view of the assumptions above, the stochastic process $[\mathbf{N}(t), M(t)]$ is a Markov process. Accounting for the packet size distribution, for queueing state $\mathbf{N}(t)=\mathbf{n}$ and background state $M(t)=m$, packets depart from queue $k$ with rate,

$$
\mu_{k}(\mathbf{n}, m)=\theta_{k} \Psi_{k}\left(\mathbf{g}_{m}, \mathbf{n}\right)
$$

such that the total departure rate in state $(\mathbf{N}(t), M(t))=(\mathbf{n}, m)$ equals,

$$
\mu(\mathbf{n}, m)=\sum_{k=1}^{K} \mu_{k}(\mathbf{n}, m)
$$

For further use, we introduce some additional notation. Let $\mathcal{C}_{k}=\left\{0,1, \ldots, C_{k}\right\}$ be the set of possible queue contents of the $k$ th queue and let $\mathcal{C}=\mathcal{C}_{1} \times \ldots \times \mathcal{C}_{K}$. The state space of our Markovian queueing model is then $\mathcal{C} \times \mathcal{M} ; S$ denotes the size of the state space $\mathcal{C} \times \mathcal{M}$. Also, $\mathbf{c}=\left[C_{1}, \ldots, C_{K}\right]$ corresponds to the case where all buffers are full; $\mathcal{M}_{\mathbf{c}}=\{[\mathbf{c}, j], j \in \mathcal{M}\}$ denotes the corresponding subset of the state space. We define $\mathbf{e}_{k}$ as the row vector of length $K$ with its $k$ th element set to 1 and all other elements equal to zero and define $\mathbf{e}$ as the row vector of ones. Finally, we introduce the global arrival rate $\lambda$ and global service rate $\mu$ which allow to simultaneously scale all the arrival rates or all service rates.

Remark 9. At the level of abstraction of the queueing model at hand, we did not specify any technological assumptions on the AP under consideration. The model at hand allows to assess the performance of the buffer behaviour at the AP for wireless systems with opportunistic scheduling like cognitive radio, micro-cell networks, Wi-Fi or WiMAX networks, and for different configurations of MU MIMO with a single AP. In particular, the present study allows for simultaneous transmissions to multiple users, while the modelling assumptions are sufficiently versatile to capture a variety of channel- and buffer-aware policies that base their scheduling decisions on the current state of the system and transmission environment.

### 4.3 Performance analysis

Having specified the modelling assumptions, we now present the numerical analysis technique. We first introduce the balance equations of the Markov chain under consideration. Expanding the stationary distribution of the Markov chain around $\mu=0$ and $\lambda=0$, with $\mu$ the global service rate and $\lambda$ the global arrival rate, we then derive approximations for the stationary distribution and various performance measures in the light-traffic and overload regime, respectively.

### 4.3.1 Balance equations

In view of the modelling assumptions introduced above, the state of the system is described by the vector $(\mathbf{n}, j)$ where the vector $\mathbf{n}=\left[n_{1}, \ldots, n_{K}\right]$ collects the states of the queues - $n_{k}$ denotes the number of packets in the $k$ th queue - and with $j \in \mathcal{M}$ the state of the background process. Moreover, let $\pi(\mathbf{n}, j)=\lim _{t \rightarrow \infty} \mathrm{P}[\mathbf{N}(t)=\mathbf{n}, M(t)=j]$ be the steady-state probability to be in state $(\mathbf{n}, j)$. As there are neither simultaneous arrivals nor departures, we find the following set of balance equations,

$$
\begin{align*}
\pi(\mathbf{n}, j)\left(\sum _ { k = 1 } ^ { K } \left(\lambda_{k}\right.\right. & \left.\left.1_{\left\{n_{k}<C_{k}\right\}}+\mu_{k}(\mathbf{n}, j) 1_{\left\{n_{k}>0\right\}}\right)+\sum_{i \in \mathcal{M} \backslash\{j\}} \alpha_{j i}\right) \\
& =\sum_{k=1}^{K} \pi\left(\mathbf{n}+\mathbf{e}_{k}, j\right) \mu_{k}\left(\mathbf{n}+\mathbf{e}_{k}, j\right) 1_{\left\{n_{k}<C_{k}\right\}} \\
& +\sum_{k=1}^{K} \pi\left(\mathbf{n}-\mathbf{e}_{k}, j\right) \lambda_{k} 1_{\left\{n_{k}>0\right\}}+\sum_{i \in \mathcal{M} \backslash\{j\}} \pi(\mathbf{n}, i) \alpha_{i j} \tag{4.1}
\end{align*}
$$

for $\mathbf{n} \in \mathcal{C}$ and $j \in \mathcal{M}$. Here $1_{\{\cdot\}}$ is the indicator which evaluates to one if its argument is true and to zero if this is not the case. State transitions correspond to arrivals at and departures from the different queues, or to state transitions of the channel. For ease of notation, we group the stationary probabilities for a given queueing state into vectors $\boldsymbol{\pi}(\mathbf{n})=[\pi(\mathbf{n}, j)]_{j \in \mathcal{M}}$. We can then rewrite the balance equations as follows,

$$
\begin{align*}
& \boldsymbol{\pi}(\mathbf{n})\left(\sum_{k=1}^{K}\left(\lambda_{k} 1_{\left\{n_{k}<C_{k}\right\}} I_{M}+M_{k}(\mathbf{n}) 1_{\left\{n_{k}>0\right\}}\right)-A\right) \\
& =\sum_{k=1}^{K} \boldsymbol{\pi}\left(\mathbf{n}+\mathbf{e}_{k}\right) M_{k}\left(\mathbf{n}+\mathbf{e}_{k}\right) 1_{\left\{n_{k}<C_{k}\right\}}+\sum_{k=1}^{K} \boldsymbol{\pi}\left(\mathbf{n}-\mathbf{e}_{k}\right) \lambda_{k} 1_{\left\{n_{k}>0\right\}} \tag{4.2}
\end{align*}
$$

with $M_{k}(\mathbf{n})$ the $M \times M$ diagonal matrix with diagonal elements $\mu_{k}(\mathbf{n}, j)$, with $I_{M}$ the $M \times M$ identity matrix and with $A$ the generator matrix of $M(t)$.

### 4.3.2 Regular perturbation

In the following subsections it is shown that a series expansion approach allows for evaluating the performance of the system under either light-traffic or overload conditions. In particular, it is shown that the series expansion of the stationary solution of the Markov process is regular (i.e., we have a regular perturbation) [29, 31, 41] and that the computational complexity of calculating the consecutive terms in the series expansion is far better than the computational complexity of
calculating the stationary distribution directly. Prior to introducing the equations for the system at hand, we outline the main ideas of the methodology.

The system of equations (4.1) takes the generic form

$$
\begin{equation*}
\boldsymbol{\pi} Q=0 \tag{4.3}
\end{equation*}
$$

where $\boldsymbol{\pi}$ is a vector which collects all stationary probabilities $\pi(\mathbf{n}, j)$ and where $Q$ is a known generator matrix whose off-diagonal elements are the transition rates between states. The row sums of the generator matrix are zero, and the matrix has negative diagonal elements and non-negative off-diagonal elements. Assume now that the entries of the generator matrix are affine functions of a system parameter $\epsilon$. In the following sections, this parameter will be the global arrival rate $\lambda$ for the light-traffic approximation and the global service rate $\mu$ for the overload approximation. As the entries of the generator matrix are affine functions of $\epsilon$, the generic equation (4.3) can be written as

$$
\begin{equation*}
\boldsymbol{\pi}^{(\epsilon)} Q=\boldsymbol{\pi}^{(\epsilon)}\left(Q^{(\mathbf{0})}+\epsilon Q^{(\mathbf{1})}\right)=0 \tag{4.4}
\end{equation*}
$$

Here we have made the dependence of the stationary solution $\pi$ on $\epsilon$ explicit. Moreover, note that $Q^{(\mathbf{0})}$ is a proper generator matrix: this is the generator matrix of the system for $\epsilon=0$. Now, assume that this Markov process is a uni-chain (the Markov process has at most one ergodic class). In this case, $\boldsymbol{\pi}^{(0)} Q^{(\mathbf{0})}=0$ has a unique normalised solution. Moreover, by Cramer's rule, one easily finds that $\boldsymbol{\pi}^{(\epsilon)}$ is an analytic function of $\epsilon$ in an open interval around $\epsilon=0$. Therefore, let $\pi_{i}$ be the $i$ th term in the series expansion of $\boldsymbol{\pi}^{(\epsilon)}$,

$$
\begin{equation*}
\boldsymbol{\pi}^{(\epsilon)}=\sum_{i=0}^{\infty} \boldsymbol{\pi}_{i} \epsilon^{i} \tag{4.5}
\end{equation*}
$$

Plugging the series expansion (4.5) into 4.4 and identifying equal powers of $\epsilon$, we get

$$
\begin{equation*}
\boldsymbol{\pi}_{0} Q^{(\mathbf{0})}=0, \quad \boldsymbol{\pi}_{i+1} Q^{(\mathbf{0})}=-\pi_{i} Q^{(\mathbf{1})} \tag{4.6}
\end{equation*}
$$

Complementing the former set of equations with the normalisation condition,

$$
\begin{equation*}
\boldsymbol{\pi}_{0} \mathbf{e}^{\prime}=1, \quad \boldsymbol{\pi}_{i} \mathbf{e}^{\prime}=0 \tag{4.7}
\end{equation*}
$$

for $i>0$, allows for recursively calculating the terms of the series expansion.
For a generic matrix $Q^{(\mathbf{0})}$, there is no gain in computational complexity as one still needs to invert this matrix while solving for the next term in the series expansion. However, for the queueing system at hand, $Q^{(0)}$ has additional structure. Indeed, for the light-traffic approximation, non- $\lambda$ transitions are either departures or changes of the channel state. Assuming a proper ordering of the states of the Markov process, the generator matrix $Q^{(\mathbf{0})}$ is block upper-diagonal, the blocks
having the size of the state space of the channel. For the overload approximation, non- $\mu$ transitions are either arrivals or changes of the channel state and - with a proper ordering of the state space - a similar block upper-diagonal structure is obtained. In either case, recursively solving the systems of equations 4.7) is considerably less involved. The equation

$$
\boldsymbol{\pi}_{0} Q^{(\mathbf{0})}=0
$$

reduces to a system of $M$ equations of $M$ unknowns, while for each $i$ the unknowns in the system

$$
\boldsymbol{\pi}_{i+1} Q^{(\mathbf{0})}=-\boldsymbol{\pi}_{i} Q^{(\mathbf{1})}
$$

can be solved in blocks of $M$ unknowns at a time due to the block upper-diagonal structure of $Q^{(\mathbf{0})}$.

### 4.3.3 Overload-traffic analysis

We first consider the balance equation for $\mu \rightarrow 0$. In particular we express the service rates as

$$
\mu_{k}(\mathbf{n}, j)=\mu \tilde{\mu}_{k}(\mathbf{n}, j)
$$

and consider the Maclaurin series expansion in $\mu$ of the steady-state probabilities:

$$
\begin{equation*}
\boldsymbol{\pi}(\mathbf{n})=\sum_{i=0}^{\infty} \boldsymbol{\pi}_{i}(\mathbf{n}) \mu^{i} . \tag{4.8}
\end{equation*}
$$

For ease of notation, let $\tilde{M}_{k}(\mathbf{n})=\mu^{-1} M_{k}(\mathbf{n})$. Note that $\tilde{M}_{k}(\mathbf{n})$ does not depend on $\mu$. Plugging the former expression into equation (4.2) and comparing terms in $\mu^{i}$, we can express the $i$ th order terms in terms of $(i-1)$ st and $i$ th order terms as follows,

$$
\begin{align*}
& \boldsymbol{\pi}_{i}(\mathbf{n}) \sum_{k=1}^{K} \lambda_{k} 1_{\left\{n_{k}<C_{k}\right\}}-\boldsymbol{\pi}_{i}(\mathbf{n}) A= \\
& 1_{\{i>0\}} \sum_{k=1}^{K} \boldsymbol{\pi}_{i-1}\left(\mathbf{n}+\mathbf{e}_{k}\right) \tilde{M}_{k}\left(\mathbf{n}+\mathbf{e}_{k}\right) 1_{\left\{n_{k}<C_{k}\right\}} \\
& \quad-1_{\{i>0\}} \boldsymbol{\pi}_{i-1}(\mathbf{n}) \sum_{k=1}^{K} \tilde{M}_{k}(\mathbf{n}) 1_{\left\{n_{k}>0\right\}} \\
&  \tag{4.9}\\
& \quad+\sum_{k=1}^{K} \boldsymbol{\pi}_{i}\left(\mathbf{n}-\mathbf{e}_{k}\right) \lambda_{k} 1_{\left\{n_{k}>0\right\}}
\end{align*}
$$

Plugging $\mathbf{n}=\mathbf{0}=[0,0, \ldots, 0]$ and $i=0$ into the former equation and postmultiplying with $\mathbf{e}^{\prime}$ leads to

$$
\begin{equation*}
\boldsymbol{\pi}_{0}(\mathbf{0}) \mathbf{e}^{\prime}=0 \tag{4.10}
\end{equation*}
$$

which implies $\boldsymbol{\pi}_{0}(\mathbf{0})=\mathbf{0}$ as the elements of $\boldsymbol{\pi}_{0}(\mathbf{0})$ are non-negative. Using the same arguments, one then shows by iteration that for all $\mathbf{n} \in \mathcal{C} \backslash\{\mathbf{c}\}$, we have $\boldsymbol{\pi}_{0}(\mathbf{n})=\mathbf{0}$ and $\boldsymbol{\pi}_{0}(\mathbf{c}) A=0$. Together with the normalisation condition $\sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}_{0}(\mathbf{n}) \mathbf{e}^{\prime}=1$, this shows that $\boldsymbol{\pi}_{0}(\mathbf{c})=\mathbf{a}$, the steady-state solution of the Markov process $M(t)$ (i.e., the normalised solution of $\mathbf{a} A=\mathbf{0}$ ).

For the higher-order terms $(i>0)$, we have

$$
\begin{array}{r}
\boldsymbol{\pi}_{i}(\mathbf{n})\left(\sum_{k=1}^{K} \lambda_{k}\right. \\
\left.1_{\left\{n_{k}<C_{k}\right\}} I_{M}-A\right)=\sum_{k=1}^{K} \boldsymbol{\pi}_{i-1}\left(\mathbf{n}+\mathbf{e}_{k}\right) \tilde{M}_{k}\left(\mathbf{n}+\mathbf{e}_{k}\right) 1_{\left\{n_{k}<C_{k}\right\}}  \tag{4.11}\\
\\
+\sum_{k=1}^{K}\left(\boldsymbol{\pi}_{i}\left(\mathbf{n}-\mathbf{e}_{k}\right) \lambda_{k}-\boldsymbol{\pi}_{i-1}(\mathbf{n}) \tilde{M}_{k}(\mathbf{n})\right) 1_{\left\{n_{k}>0\right\}}
\end{array}
$$

For $\mathbf{n} \neq \mathbf{c}$, the matrix on the left-hand side is invertible. Hence, we can calculate the probabilities $\boldsymbol{\pi}_{i}(\mathbf{n})$ in lexicographical order. For $\mathbf{n}=\mathbf{c}$, we get

$$
\begin{equation*}
\boldsymbol{\pi}_{i}(\mathbf{c}) A=\sum_{k=1}^{K}\left(-\boldsymbol{\pi}_{i}\left(\mathbf{c}-\mathbf{e}_{k}\right) \lambda_{k}+\boldsymbol{\pi}_{i-1}(\mathbf{c}) \tilde{M}_{k}(\mathbf{c})\right) \tag{4.12}
\end{equation*}
$$

and the matrix on the left-hand side is not invertible. A solution of this equation takes the form

$$
\begin{equation*}
\boldsymbol{\pi}_{i}(\mathbf{c})=\sum_{k=1}^{K}\left(-\boldsymbol{\pi}_{i}\left(\mathbf{c}-\mathbf{e}_{k}\right) \lambda_{k}+\boldsymbol{\pi}_{i-1}(\mathbf{c}) \tilde{M}_{k}(\mathbf{c})\right) A^{\#}+\kappa_{i} \mathbf{a} \tag{4.13}
\end{equation*}
$$

for any $\kappa_{i}$. Here, $A^{\#}=\left(A+\mathbf{e}^{\prime} \mathbf{a}\right)^{-1}-\mathbf{e}^{\prime} \mathbf{a}$ is the group inverse of $A$. Finally, the remaining unknown $\kappa_{i}$ follows from the normalisation condition

$$
\begin{equation*}
\sum_{\mathbf{n} \in \mathcal{C}} \pi_{i}(\mathbf{n}) \mathbf{e}^{\prime}=0 \tag{4.14}
\end{equation*}
$$

In view of the calculations above, one easily verifies that the numerical complexity of the algorithm for the $L$ th order expansion is $O\left(L M^{2} S\right)$ as there are $S / M$ blocks, $L$ terms in the recursion and the operations with blocks have complexity $O\left(M^{3}\right)$.

### 4.3.4 Light-traffic analysis

Similar arguments can be developed for the case of light-traffic conditions, that is, we set $\lambda_{k}=\lambda \tilde{\lambda}_{k}$ and consider an expansion of the form

$$
\boldsymbol{\pi}(\mathbf{n})=\sum_{i=0}^{\infty} \boldsymbol{\pi}_{i}(\mathbf{n}) \lambda^{i}
$$

In view of the balance equations, the terms of this series expansion adhere

$$
\begin{align*}
& \boldsymbol{\pi}_{i}(\mathbf{n})\left(\sum_{k=1}^{K} M_{k}(\mathbf{n}) 1_{\left\{n_{k}>0\right\}}-A\right)=\sum_{k=1}^{K} \boldsymbol{\pi}_{i}\left(\mathbf{n}+\mathbf{e}_{k}\right) M_{k}\left(\mathbf{n}+\mathbf{e}_{k}\right) 1_{\left\{n_{k}<C_{k}\right\}} \\
- & 1_{\{i>0\}} \boldsymbol{\pi}_{i-1}(\mathbf{n}) \sum_{k=1}^{K} \tilde{\lambda}_{k} 1_{\left\{n_{k}<C_{k}\right\}}+1_{\{i>0\}} \sum_{k=1}^{K} \boldsymbol{\pi}_{i-1}\left(\mathbf{n}-\mathbf{e}_{k}\right) \tilde{\lambda}_{k} 1_{\left\{n_{k}>0\right\}} \tag{4.15}
\end{align*}
$$

For $i=0$, we can show that $\boldsymbol{\pi}_{0}(\mathbf{n})=\mathbf{0}$ for $\mathbf{n} \neq \mathbf{0}$ and $\boldsymbol{\pi}_{0}(\mathbf{0})=\mathbf{a}$. For $i>0$ and $\mathbf{n} \neq \mathbf{0}$, we can recursively calculate all $\boldsymbol{\pi}_{i}(\mathbf{n})$ in reverse lexicographical order as the matrix on the left-hand side is invertible. For $\mathbf{n}=\mathbf{0}$, we get

$$
\begin{equation*}
\left.\boldsymbol{\pi}_{i}(\mathbf{0})=-\sum_{k=1}^{K} \boldsymbol{\pi}_{i}\left(\mathbf{e}_{k}\right) M_{k}\left(\mathbf{e}_{k}\right)+\boldsymbol{\pi}_{i-1}(\mathbf{0}) \sum_{k=1}^{K} \tilde{\lambda}_{k}\right) A^{\#}+\tilde{\kappa}_{i} \mathbf{a} \tag{4.16}
\end{equation*}
$$

where $\tilde{\kappa}_{i}$ can be determined from the normalisation condition (4.14).

### 4.3.5 Performance metrics

In order to quantify the performance of the system at hand we consider the following metrics for the $k$ th queue: the mean queue content $\mathrm{E}\left[Q_{k}\right]$, the variance of the queue content $\operatorname{var}\left[Q_{k}\right]$ and the blocking probability $b_{k}$. As Poisson arrivals see time averages, the blocking probability for the $k$ th queue is the probability that the number of packets in this queue is equal to the queue capacity, $b_{k}=\mathrm{P}\left[n_{k}=C_{k}\right]$. Additionally, we calculate the mean of the total system content $\mathrm{E}\left[Q^{\text {total }}\right]$. These characteristics can be expressed in terms of the stationary distribution $\boldsymbol{\pi}(\mathbf{n})$. We then find approximations of these metrics by replacing the stationary distribution by its $L$ th order expansion,

$$
\begin{align*}
\mathrm{E}\left[Q^{\text {total }}\right] & =\sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime} \mathbf{n} \mathbf{e}^{\prime} \approx \sum_{i=0}^{L} \sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}_{i}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime} \mathbf{n} \mathbf{e}^{\prime} \epsilon^{i}, \\
\mathrm{E}\left[Q_{k}\right] & =\sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime} \mathbf{n e}{ }_{k}^{\prime} \approx \sum_{i=0}^{L} \sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}_{i}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime} \mathbf{n e}_{k}^{\prime} \epsilon^{i}, \\
b_{k} & =\sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime} 1_{\left\{\mathbf{n e}_{k}^{\prime}=C_{k}\right\}} \approx \sum_{i=0}^{L} \sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}_{i}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime} \epsilon^{i} 1_{\left\{\mathbf{n} \mathbf{e}_{k}^{\prime}=C_{k}\right\}}, \\
\operatorname{var}\left[Q_{k}\right] & =\sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime}\left(\mathbf{n e}_{k}^{\prime}\right)^{2}-\mathrm{E}\left[Q_{k}\right]^{2} \\
& \approx \sum_{i=0}^{L} \sum_{\mathbf{n} \in \mathcal{C}} \boldsymbol{\pi}_{i}^{(\epsilon)}(\mathbf{n}) \mathbf{e}^{\prime}\left(\mathbf{n} \mathbf{n}_{k}^{\prime}\right)^{2} \epsilon^{i}-\mathrm{E}\left[Q_{k}\right]^{2}, \tag{4.17}
\end{align*}
$$

where the dependence on $\epsilon$ has again been made explicit, $\epsilon$ being the service rate $\mu$ in the overload expansion or the arrival rate $\lambda$ for the light-traffic case.

### 4.4 Modelling the wireless environment

In this section, we introduce a particular channel model for the multi-channel wireless communication scenario at hand. For a wide range of applications, fluctuations in a wireless communication link can be efficiently modelled by means of Markov chains. For a single channel, multiple authors have proposed Markovian models before, see e.g. [27, 42-44]. We here adapt the discrete-time channel model of [27] to the continuous-time Markov chain setting. As we also investigate the influence of cross-channel correlation, we extend the single-channel approach of [27] to multiple correlated channels. The proposed approach for both singleand multi-channel models also extends to other channel models including Rician fading channels and Nakagami- $m$ fading channels [18].

### 4.4.1 Single-channel model

To model the behaviour of a single Rayleigh fading channel by a Markov chain with state space $\mathcal{J}=\{1,2, \ldots, J\}$, we construct a Markov chain such that the stationary distribution of the Markov model closely matches the distribution of the Signal-to-Noise ratio (SNR), as well as the rates in which certain levels are crossed.

For a Rayleigh fading channel, the stationary distribution of the SNR is exponentially distributed,

$$
\mathrm{P}[\mathrm{SNR} \leq \eta]=F(\eta)=1-\exp \left(-\frac{\eta}{\nu}\right)
$$

where $\nu$ denotes the mean SNR. To closely match this distribution, we first discretise the distribution. That is, we choose levels $\eta_{0}<\eta_{1}<\eta_{2}<\ldots<\eta_{J}$, and define,
$y_{j}=\mathrm{P}\left[\eta_{j-1}<\mathrm{SNR} \leq \eta_{j}\right]=F\left(\eta_{j}\right)-F\left(\eta_{j-1}\right)=\exp \left(-\frac{\eta_{j-1}}{\nu}\right)-\exp \left(-\frac{\eta_{j}}{\nu}\right)$.
That is, $y_{j}$ is the probability that the SNR lies in the interval $\left(\eta_{j-1}, \eta_{j}\right]$.
The time variations of the SNR are characterised by the Doppler frequency effect, which is caused by the motion of the mobile nodes. In particular, let $N_{j}$ denote the mean number of times per time unit the SNR crosses the threshold $\eta_{j}$
downward. Obviously, $N_{j}$ is also the mean number of times per time unit the SNR crosses the threshold $\eta_{j}$ upward. $N_{j}$ is given by [45],

$$
\begin{equation*}
N_{j}=\sqrt{\frac{2 \pi \eta_{j}}{\nu}} f \exp \left(-\frac{\eta_{j}}{\nu}\right) \tag{4.19}
\end{equation*}
$$

where $f$ denotes the Doppler frequency and where $\nu$ is the mean SNR as before.
We now construct a Markov chain $Y(t)$ with stationary distribution $\left\{y_{j}, j \in\right.$ $\mathcal{J}\}$ such that the mean number of transitions per time unit from state $j$ to state $j+1$ equals $N_{j}$. Indeed, a transition from state $j$ to state $j+1$ corresponds to crossing threshold $\eta_{j}$. Downward crossings are defined likewise. This leads to the following transition rates for $j, k \in \mathcal{J}$,

$$
a_{j, k}= \begin{cases}\frac{N_{j}}{y_{j}} & \text { for } k=j+1 ; \\ \frac{N_{j-1}}{y_{j}} & \text { for } k=j-1 ; \\ 0 & \text { otherwise }\end{cases}
$$

It is now easily verified by checking the local balance equations that $\left\{y_{j}, j \in \mathcal{J}\right\}$ is the stationary distribution of the Markov chain with transition rates $\left\{a_{j, k}, j, k \in\right.$ $\mathcal{J}\}$. Hence we have found a Markov process which agrees with the (discretised) stationary distribution, as well as with the level crossings specified by the Doppler effect.

Once the generator matrix $\tilde{A}$ with transition rates $a_{j, k}$ is obtained for a single channel as above, we can easily expand the model for the case of multiple independent channels. We can do so by merging $K$ single-channel models, the $k$ th model having state space $\mathcal{J}_{k}=\left\{1, \ldots, J_{k}\right\}$ into a multidimensional Markov process. The state space of the joint model is then $\mathcal{M}=\mathcal{J}_{1} \times \ldots \mathcal{J}_{K}$. In other words, the state of the channels is described by a vector, the $k$ th element in this vector denoting the state of the $k$ th channel. The generator matrix of this Markov chain is,

$$
\begin{equation*}
A=\bigoplus_{k=1}^{K} A_{k} \tag{4.20}
\end{equation*}
$$

where $A_{k}$ is the $J_{k} \times J_{k}$ generator matrix of the $k$ th channel and where $\oplus$ denotes the Kronecker sum.

### 4.4.2 Multiple channels

In order to assess the effects of channel correlations, we now construct a channel model where (i) the characteristics of the single-channel models are given and (ii) some form of correlation between the channels is introduced. Moreover, our channel model with cross-correlation will have the same state space as a similar model without correlation.

Copulas We consider a model with $K$ Rayleigh fading channels. Let $\nu_{k}$ denote the mean SNR of the $k$ th channel and let $f_{k}$ denote the Doppler frequency of the $k$ th channel. Given the distribution $F_{k}$ of the SNR of the $k$ th channel for $k=1, \ldots, K$, we construct a joint distribution of the channel SNRs by means of a copula. A $K$-dimensional copula $C(\mathbf{x}), \mathbf{x} \in[0,1]^{K}$, is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform, see [46]. Hence, for a copula one has,

$$
C(1, \ldots, 1, u, 1, \ldots, 1)=u
$$

for $0 \leq u \leq 1$. Given a copula $C_{\theta}$ with parameter $\theta$, the $K$-dimensional distribution

$$
F(\mathbf{x}, \theta)=C_{\theta}\left(F_{1}\left(x_{1}\right), \ldots, F_{K}\left(x_{K}\right)\right)
$$

then has the required marginal distributions. The parameter $\theta$ can then be used to introduce the required amount of cross-channel correlation. For the numerical examples, we adopt the Clayton and Vine copulas.

The Clayton copula is an Archimedean copula with generator $\phi(x)=\max ((1+$ $\theta x)^{-1 / \theta}, 0$, yielding the copula,

$$
C_{\theta}(\mathbf{x})=\max \left(\sum_{k=1}^{K} x_{k}^{-\theta}-(K-1), 0\right)^{-1 / \theta}
$$

for $\mathbf{x}=\left[x_{1}, \ldots, x_{K}\right] \in\left(\mathbb{R}^{+}\right)^{K}$ and $\theta \in[-1 /(d-1), \infty) \backslash\{0\}$. For the Rayleigh fading channels we therefore get,

$$
F(\mathbf{x}, \theta)=\max \left(\sum_{k=1}^{K}\left(1-\exp \left(-x_{k} / \nu_{k}\right)\right)^{-\theta}-(K-1), 0\right)^{-1 / \theta}
$$

By construct, the Clayton copula is symmetric. To allow for a richer correlation structure, one can construct a Vine copula. Vine copulas combine multiple bivariate copulas into a single multivariate copula. Such a pair-copula decomposition allows for a flexible and intuitive way of extending bivariate copulas to higher dimensions [46-48]. For the general theory on Vine copulas, we refer to [48]. We will apply a 3-dimensional regular D-Vine in the remainder. The density of the multivariate distribution is then given by,

$$
\begin{aligned}
f\left(x_{1}, x_{2}, x_{3}\right)=f_{1}( & \left.x_{1}\right) f_{2}\left(x_{2}\right) f_{3}\left(x_{3}\right) c_{12}\left(F_{1}\left(x_{1}\right), F_{2}\left(x_{2}\right)\right) \\
& \times c_{23}\left(F_{2}\left(x_{2}\right), F_{3}\left(x_{3}\right)\right) c_{13 \mid 2}\left(F_{1 \mid 3}\left(x_{1} \mid x_{2}\right), F_{3 \mid 2}\left(x_{3} \mid x_{2}\right)\right)
\end{aligned}
$$

or, additionally assuming conditional independence, by,

$$
f\left(x_{1}, x_{2}, x_{3}\right)=f_{1}\left(x_{1}\right) f_{2}\left(x_{2}\right) f_{3}\left(x_{3}\right) c_{12}\left(F_{1}\left(x_{1}\right), F_{2}\left(x_{2}\right)\right)
$$

$$
\begin{equation*}
\times c_{23}\left(F_{2}\left(x_{2}\right), F_{3}\left(x_{3}\right)\right) c_{13 \mid 2}\left(F_{1}\left(x_{1}\right), F_{3}\left(x_{3}\right)\right) . \tag{4.21}
\end{equation*}
$$

Here each $c_{*}(\cdot, \cdot)$ is a bivariate copula, $F_{k}(\cdot)$ is the univariate distribution of the SNR of the $k$ th channel as above, and $f_{k}(\cdot)$ is the corresponding density function.

Discretisation We now again discretise the SNRs. Assuming we discretise the SNR of the $k$ th channel in $J_{k}$ levels, let $\eta_{k, 0}<\eta_{k, 1}<\eta_{k, 2}<\ldots<\eta_{k, J_{k}}$ denote the thresholds of these SNR levels. We then introduce the multivariate discrete distribution,

$$
y_{\mathbf{j}}=\mathrm{P}\left[\eta_{1, j_{1}-1}<\mathrm{SNR}_{1} \leq \eta_{1, j_{1}}, \ldots, \eta_{K, j_{K}-1}<\mathrm{SNR}_{K} \leq \eta_{K, j_{K}}\right]
$$

with $\mathbf{j}=\left[j_{1}, \ldots, j_{K}\right] \in \mathcal{M}=\mathcal{J}_{1} \times \ldots \times \mathcal{J}_{K}$, with $\mathcal{J}_{k}=\left\{1, \ldots, J_{k}\right\}$ as before, and where $\mathrm{SNR}_{k}$ denotes the SNR of the $k$ th channel. We can of course express $y_{\mathbf{j}}$ in terms of $F(\mathbf{x})$ as follows,

$$
y_{\mathbf{j}}=\sum_{\mathbf{i} \in\{0,1\}^{K}}(-1)^{\|\mathbf{i}\|} F\left(\eta_{1, j_{1}-i_{1}}, \ldots, \eta_{K, j_{K}-i_{K}}\right),
$$

with $\mathbf{i}=\left[i_{1}, \ldots, i_{K}\right]$ and where $\|\mathbf{i}\|=\sum_{k=1}^{K} i_{k}$ is the $\ell_{1}$ norm. For example, for $K=3$ we have,

$$
\begin{aligned}
y_{j_{1}, j_{2}, j_{3}}= & F\left(\eta_{1, j_{1}}, \eta_{2, j_{2}}, \eta_{3, j_{3}}\right)-F\left(\eta_{1, j_{1}-1}, \eta_{2, j_{2}}, \eta_{3, j_{3}}\right) \\
& -F\left(\eta_{1, j_{1}}, \eta_{2, j_{2}-1}, \eta_{3, j_{3}}\right)+F\left(\eta_{1, j_{1}-1}, \eta_{2, j_{2}-1}, \eta_{3, j_{3}}\right) \\
& -F\left(\eta_{1, j_{1}}, \eta_{2, j_{2}}, \eta_{3, j_{3}-1}\right)+F\left(\eta_{1, j_{1}-1}, \eta_{2, j_{2}}, \eta_{3, j_{3}-1}\right) \\
& \quad+F\left(\eta_{1, j_{1}}, \eta_{2, j_{2}-1}, \eta_{3, j_{3}-2}\right)-F\left(\eta_{1, j_{1}-1}, \eta_{2, j_{2}-1}, \eta_{3, j_{3}-1}\right) .
\end{aligned}
$$

Construction of the Markov process As for the single-channel case, the time variations of the SNR are characterised by the Doppler frequency effect, which is caused by the motion of the mobile nodes. We now construct a Markov chain such that the mean number of upcrossings of the thresholds for the $k$ th channel is in accordance with the Doppler effect. That is, the mean number of upcrossings of level $\eta_{k, j}$ by the $k$ th channel equals,

$$
N_{k, j}=\sqrt{\frac{2 \pi \eta_{k, j}}{\nu}} f_{k} \exp \left(-\frac{\eta_{k, j}}{\nu}\right) .
$$

Additionally assuming that channels do not cross at the same time, we find for $\mathbf{i}, \mathbf{j} \in \mathcal{M}$,

$$
\alpha_{\mathbf{j}, \mathbf{i}}= \begin{cases}\frac{N_{k, j_{k}}}{y_{\mathbf{k}}} & \text { for } \mathbf{i}=\mathbf{j}+\mathbf{e}_{k} ; \\ \frac{N_{k, j_{k}-1}}{y_{\mathbf{j}}} & \text { for } \mathbf{i}=\mathbf{j}-\mathbf{e}_{k} ; \\ 0 & \text { otherwise } .\end{cases}
$$

As for the single-channel case, it is now easily verified by checking the local balance equations that $\left\{y_{\mathbf{j}}, \mathbf{j} \in \mathcal{M}\right\}$ is the stationary distribution of the Markov chain with transition rates $\left\{\alpha_{\mathbf{j}, \mathbf{i}}, \mathbf{j}, \mathbf{i} \in \mathcal{M}\right\}$. Hence we have found a Markov process which agrees with the (discretised) stationary (multivariate) distribution, as well as with the level crossings specified by the Doppler effect.

### 4.4.3 Channel State Information

The Markov chains for the single and multiple channels above describe the evolution of the SNR of the channels. It now remains to map the states of these Markov chains on the transmission rates that can be achieved. We describe the mapping for the multivariate case, the single-channel model being a particular example.

The maximal achievable channel capacity $\gamma$ for a given SNR $\eta$ and bandwidth $B$, is given by the Shannon-Hartley theorem [45],

$$
\gamma=B \log _{2}(1+\eta)
$$

As the transmission rate is proportional to the channel capacity, we introduce the following mapping $\mathbf{j}=\left[j_{1}, \ldots, j_{K}\right] \rightarrow \mathbf{g}_{\mathbf{j}}$ from $\mathcal{M}$ to $[0,1]^{K}$,

$$
\mathbf{g}_{\mathbf{j}}=\left[\frac{\log _{2}\left(1+\xi_{1, j_{1}}\right)}{\log _{2}\left(1+\eta_{\text {best }}\right)}, \ldots, \frac{\log _{2}\left(1+\xi_{K, j_{K}}\right)}{\log _{2}\left(1+\eta_{\text {best }}\right)}\right]
$$

where $\eta_{\text {best }}$ is the best SNR that is achievable (for all channels) and where $\xi_{k, j}$ is the average SNR of the $k$ th channel in state $j$,

$$
\xi_{k, j}=\int_{\eta_{k, j-1}}^{\eta_{k, j}} \eta d F_{k}(\eta)
$$

The $k$ th element $g_{\mathbf{j} k}$ of the vector $\mathbf{g}_{\mathbf{j}}$ denotes the fraction of the maximal transmission rate that is available for the $k$ th channel when the channel state is $\mathbf{j}$. The mapping from SNR to rate depends on the system specifications and can vary from one transmitter to another. The value $g_{\mathbf{j} k}$ is referred to as the channel state information and is made available to the opportunistic scheduler.

### 4.4.4 Examples of schedulers

For the sake of illustration, we have implemented several schedulers including purely opportunistic, purely non-opportunistic and weighted schedulers. We describe the different schedulers in detail below.

MaxRate A first example of a greedy opportunistic scheduler is the MaxRate scheduler. This scheduler one serves the mobile nodes with the best channel conditions. Let $\kappa_{M R}(\mathbf{j})$ be the set

$$
\kappa_{M R}(\mathbf{j})=\underset{k \in\{1, \ldots, K\}}{\arg \max } g_{\mathbf{j} k},
$$

where $g_{\mathbf{j} k}$ is the channel state information for user $k$, which depends on the state of the channel $\mathbf{j}$. For MaxRate, we have the following service rate,

$$
\mu_{k}(\mathbf{n}, \mathbf{j})= \begin{cases}\mu g_{\mathbf{j} k} \frac{1}{\left|\kappa_{M R}(\mathbf{j})\right|} & \text { for } k \in \kappa_{M R}(\mathbf{j}) \\ 0 & \text { otherwise }\end{cases}
$$

Here $\left|\kappa_{M R}(\mathbf{j})\right|$ is the cardinality of $\kappa_{M R}(\mathbf{j})$. This scheduler was first considered in [4] for single-cell MU communication.

MaxWeight In contrast to MaxRate, the MaxWeight scheduler selects the user with the maximum weight, which is calculated as the product of queue length and channel quality, see [40]. For $\mathbf{N}(t)=\mathbf{n}$ and $M(t)=\mathbf{j}$, MaxWeight selects users at time $t$ from the set,

$$
\kappa_{M W}(\mathbf{n}, \mathbf{j})=\underset{k \in\{1, \ldots, K\}}{\arg \max } g_{\mathbf{j} k} n_{k}
$$

Hence this results in the following service rate for MaxWeight,

$$
\mu_{k}(\mathbf{n}, \mathbf{j})= \begin{cases}\mu g_{\mathbf{j} k} \frac{1}{\left|\kappa_{M W}(\mathbf{j})\right|} & \text { for } k \in \kappa_{M W}(\mathbf{j}) \\ 0 & \text { otherwise }\end{cases}
$$

Longest connected queue Schedulers may also not account for the channel state at all. An example of a non-opportunistic scheduler is one that chooses the longest queue. The scheduler is shown to be stable for dynamic server allocation to parallel queues with randomly varying connectivity in [39] and called the Longest Connected Queue (LCQ). For $\mathbf{N}(t)=\mathbf{n}$ and $M(t)=\mathbf{j}$, the scheduler serves mobile nodes from the set

$$
\kappa_{L C Q}=\underset{k \in\{1, \ldots, K\}}{\arg \max } n_{k}
$$

Notice that also in this case the actual service rate $\mu_{k(\mathbf{n})}(\mathbf{n}, j)$ for node $k(\mathbf{n})$ does depend on the channel condition. Indeed, we have,

$$
\mu_{k}(\mathbf{n}, \mathbf{j})= \begin{cases}\mu g_{\mathbf{j} k} \frac{1}{\left|\kappa_{L C Q}(\mathbf{j})\right|} & \text { for } k \in \kappa_{L C Q}(\mathbf{j}) \\ 0 & \text { otherwise }\end{cases}
$$

Processor sharing Finally, we mention two schedulers which are inspired by discriminatory (DPS) and generalised processor sharing (GPS) [31], but with weights set to reflect the channel conditions. The share of transmission resources assigned for each user takes one of the following forms:

$$
\text { DPS: } s_{k}(\mathbf{n}, \mathbf{j})=\frac{g_{\mathbf{j} k} n_{k}}{\sum_{\ell=1}^{K} g_{\mathbf{j} \ell} n_{\ell}}
$$

$$
\text { GPS: } s_{k}(\mathbf{n}, \mathbf{j})=\frac{g_{\mathbf{j} k} 1_{\left\{n_{k}>0\right\}}}{\sum_{\ell=1}^{K} g_{\mathbf{j} \ell} 1_{\left\{n_{\ell}>0\right\}}},
$$

where $g_{\mathbf{j} k}$ is the channel quality for user $k$ when the channel state is $\mathbf{j}$. Eventually, the transmission rate of the $k$ th mobile user $\mu_{k}(\mathbf{n}$,$) depends on the overall system$ transmission rate $\mu$, the share of transmission resources $s_{k}(\mathbf{n}, \mathbf{j})$ assigned to user $k$ and the quality of the radio link $g_{\mathrm{j} k}$ :

$$
\mu_{k}(\mathbf{n}, \mathbf{j})=\mu g_{\mathbf{j} k} s_{k}(\mathbf{n}, \mathbf{j}) .
$$

### 4.5 Results

In this section we compare various schedulers by some numerical examples. We consecutively consider systems with independent channels and systems with crosschannel correlation. The following performance metrics are calculated by means of 4.17): the mean queue content $\mathrm{E}\left[Q_{k}\right]$, the variance of the queue content $\operatorname{var}\left[Q_{k}\right]$ and the blocking probability $b_{k}$ of the $k$ th queue. Assuming identical queues and channels, we provide results for a single queue, and therefore drop the index $k$ in the figures. As our analysis is an approximation, we validate our results by simulations. The simulation results are obtained by means of the standard Gillespie algorithm, see for example [49].

### 4.5.1 Independent channels

We first assume that all channels are independent and identical with Doppler frequency $f_{k}=100 \mathrm{~Hz}$ and mean SNR level $\nu_{k}=30 \mathrm{~dB}$ for $k=1, \ldots, K$. To construct the channel model, we partition the SNR into 3 ranges, the thresholds being chosen as follows: $\eta_{0}=0 \mathrm{~dB}, \eta_{1}=20 \mathrm{~dB}, \eta_{2}=30 \mathrm{~dB}, \eta_{3}=40 \mathrm{~dB}$. Note that various strategies of SNR partitioning can be applied in order to establish SNR thresholds [50], the choice made here mainly being for illustration purposes. With these parameters, the method of channel modelling described in subsection 4.4.1 results in the following transition rate matrix:

$$
A_{k}=\left(\begin{array}{ccc}
-75.369 & 75.369 & 0 \\
13.357 & -30.530 & 17.173 \\
0 & 25.069 & -25.069
\end{array}\right)
$$

In the figures below, we consider systems with $K=4$ mobile users experiencing independent and stochastically identical channel conditions. Up to $C=10$ packets can be stored at the buffer of the AP for each mobile user; packets arriving at a full buffer are rejected. We assume a maximum achievable downlink transmission rate of $10 \mathrm{Mbit} / \mathrm{s}$ and an average packet size of 1.125 MB . Note that with the above assumptions, the channel process with transition rate matrix $A=\oplus_{k=1}^{4} A_{k}$
has $3^{4}$ states, whereas the number of queueing states amounts to $11^{4}$ states. Hence, the Markov process at hand has 1.185 .921 states.

In a first set of figures, we compare the MaxRate, MaxWeight and LCQ schedulers. Figures 4.2 and 4.3 show the mean and variance of the queue content as well as the blocking probabilities for these schedulers, in the light-traffic and in the overloaded regime, respectively. For each of these performance measures, we plot the 5th, 15 th and 50th order expansions, as well as simulation results for comparison. Both simulation and numerical results indicate that the purely opportunistic MaxRate scheduler performs better than MaxWeight and LCQ in both the light-traffic regime and the overloaded regime. The reason for this advantage can be explained as follows: MaxRate always serves a customer with the maximum available transmission rate, and therefore, maximises the throughput of the system. In contrast, the MaxWeight policy selects the user with the highest product of queue length and transmission rate. Under this strategy the system benefits from the multi-channel diversity while preventing service starvation in the queues with constantly poor channel quality. Fairness of the MaxWeight scheduler, however, comes at the cost of a decrease of the system throughput as compared to the purely opportunistic MaxRate. In comparison with LCQ, MaxWeight performs slightly better than the LCQ scheduler, due to the utilisation of the channel information. Recall that LCQ does not benefit from the channel information and always serves the node with the longest queue, even if the channel conditions are unfavourable for transmission. This approach guarantees fairness, however, impairs the system performance in a transmission environment with a high MU diversity. Numerical results show that although MaxWeight performs better than LCQ, the performance metrics are rather close to each other.

We now assess the accuracy of the series expansions. Obviously, by increasing the order of the expansion, the approximations improve, as they are known to be converging to an exact match within the region of convergence of the expansion. While the 5 th order expansion is only accurate in a small region around $\lambda=0$ or $\mu=0$, the 15 th order approximation is already accurate in a much wider region, while the 50th order expansion is accurate in an even wider region. Note that the series expansions of all performance measures (for a particular scheduler) have the same region of convergence, as they are derived from the same expansion of the steady-state distribution. In fact, the higher order expansions allow for heuristically determining the region of convergence as we get a very accurate match in the region of convergence, followed by an almost immediate and extensive deviation of the true value. This is the case for the 50th order expansion, higher order expansions (not shown here) further confirming that the position of this sudden deviation is fixed. This then indeed strongly suggests that this position corresponds with the boundary of the region of convergence. On the figures, the limits of the regions of convergence of the light traffic approximation are $\lambda \in[0,10.1] \mathrm{Mbit} / \mathrm{s}$


Figure 4.2: Mean queue content, variance and blocking probability in light traffic regime for systems with $K=4$ mobile users and $M=3$ channel states, working under MaxRate(MR), MaxWeight(MW) and LCQ scheduling disciplines


Figure 4.3: Mean queue content, variance and blocking probability in overload traffic regime for systems with $K=4$ mobile users and $M=3$ channel states, working under MaxRate(MR), MaxWeight(MW) and LCQ scheduling disciplines.
for MaxRate, $\lambda \in[0,6.3]$ Mbit/s for MaxWeight, and $\lambda \in[0,6.8] \mathrm{Mbit} / \mathrm{s}$ for LCQ. For the overload expansions in Figure 4.3, we observe the convergence regions $\mu \in[0,8.6] \mathrm{Mbit} / \mathrm{s}, \mu \in[0,7.9] \mathrm{Mbit} / \mathrm{s}$ and $\mu \in[0,4.7] \mathrm{Mbit} / \mathrm{s}$ for the MaxRate, MaxWeight and LCQ schedulers respectively.

As the blocking probability decreases almost linearly for increasing $\mu$, the 5th order expansion seems rather accurate in a wide region. The match beyond the region of convergence for low order expansions is coincidence and does not further improve by increasing the order of the expansion. Indeed, the higher order expansions reveal the limits of the region of convergence. For the light-traffic approximations, the blocking probability is very small in the region of convergence for the MaxWeight and LCQ schedulers; the values for which the blocking probability starts increasing lie outside the region of convergence. In contrast, for the MaxRate scheduler, we do find a good match in the region where the blocking probabilities start increasing (the interval $\lambda \in[7,9.5]$ Mbit).

The next set of figures compare the DPS and GPS schedulers. Figures 4.4 and 4.5 depict the mean and variance of the queue content and the blocking probability vs. the global arrival rate and vs. the global service rate, respectively. DPS and GPS are opportunistic schedulers that allocate resources among the mobile users for simultaneous transmission. All users with nonempty queues receive a share of the service proportional to their channel qualities. GPS always allocates resources primarily to the users with favourable channel conditions, while DPS also takes into account the queue content of the customers. The comparison of these two schedulers is similar to the MaxRate and MaxWeight comparison. GPS is purely opportunistic and provides better throughput, while DPS guarantees fairness and prevents service starvation for users experiencing poor channel conditions. The numerical results in Figures 4.4 and 4.5 indeed confirm these findings, though the difference between GPS and DPS is not as outspoken as the difference between MaxWeight and MaxRate. Moreover, in the light-traffic regime, GPS and DPS hardly differ. This is not unexpected as DPS and GPS only differ once the queue size grows beyond 1 , which does not occur frequently in light traffic.

Regarding the accuracy of the approximations, we again observe that increasing the order of the approximations improves the accuracy, in the region of convergence of the series expansions. Compared to the previous plots (figures 4.2 and 4.3), the region of convergence is considerably wider. For DPS we have $\lambda \in$ $[0,8.3] \mathrm{Mbit} / \mathrm{s}$ and $\mu \in[0,10.9] \mathrm{Mbit} / \mathrm{s}$, while for GPS, we have $\lambda \in[0,8.4] \mathrm{Mbit} / \mathrm{s}$ and $\mu \in[0,9.7]$ Mbit/s.

### 4.5.2 Correlated channels

We now study opportunistic scheduling with correlated channels. We retain the single channel assumptions of the preceding section: the mean SNR for every


Figure 4.4: Mean queue content, variance and blocking probability in light traffic regime for systems with $K=4$ mobile users and $M=3$ channel states, working under DPS and GPS scheduling disciplines


Figure 4.5: Mean queue content, variance and blocking probability in overload traffic regime for systems with $K=4$ mobile users and $M=3$ channel states, working under DPS and GPS scheduling disciplines
channel is $\nu=30 \mathrm{~dB}$, while the Doppler frequency is $f_{k}=100 \mathrm{~Hz}$. We use the same threshold values as well: $\eta_{0}=0 \mathrm{~dB}, \eta_{1}=20, \eta_{2}=30 \mathrm{~dB}, \eta_{3}=40 \mathrm{~dB}$. The AP serves 3 mobile users, where each user can store up to $C=30$ packets. Having defined the characteristics of a single channel, we study the following three different correlated channel scenarios.

1. Using a Clayton copula with parameter $\theta=1$, we obtain the following channel correlation matrix,

$$
\mathbf{R}_{1}=\left(\begin{array}{ccc}
1 & 0.36948717 & 0.36948717  \tag{4.22}\\
0.36948717 & 1 & 0.36948717 \\
0.36948717 & 0.36948717 & 1
\end{array}\right)
$$

where the $k j$ th element of $\mathbf{R}$ denotes Pearson's correlation coefficient of the SNR of the $k$ th and $j$ th channel,

$$
r_{k \ell}=\frac{\sigma_{k \ell}^{2}}{\sqrt{\sigma_{k k}^{2} \sigma_{\ell \ell}^{2}}}
$$

with

$$
\sigma_{k \ell}^{2}=\sum_{\mathbf{j} \in \mathcal{M}} y_{\mathbf{j}} \xi_{k, j_{k}} \xi_{\ell, j_{\ell}}-\left(\sum_{\mathbf{j} \in \mathcal{M}} y_{\mathbf{j}} \xi_{k, j_{k}}\right)\left(\sum_{\mathbf{j} \in \mathcal{M}} y_{\mathbf{j}} \xi_{\ell, j_{\ell}}\right)
$$

where we used the notation of subsection 4.4.2. For any two channels, we obtain the same correlation. This is not unexpected due to the symmetry of Archimedean copulas.
2. A Vine copula allows for non-symmetric channel correlations. We consider two vine copulas that are build using bivariate Gaussian copulas. We use expression 4.21 , where the copulas $C_{12}, C_{23}$ and $C_{13 \mid 2}$ are Gaussian copulas with correlation coefficients $\rho_{12}=0.2, \rho_{23}=0.1$ and $\rho_{13 \mid 2}=0.3$. This parameter setting results in the following correlation matrix,

$$
\mathbf{R}_{2}=\left(\begin{array}{ccc}
1 & 0.07591856 & 0.23769156  \tag{4.23}\\
0.07591856 & 1 & 0.15236588 \\
0.23769156 & 0.15236588 & 1
\end{array}\right)
$$

3. The third correlated channel is again based on a Vine copula with Gaussian bivariate copulas, but now with correlation parameters $\rho_{12}=0.8, \rho_{23}=0.7$, $\rho_{13 \mid 2}=0.6$. For this set of parameters, we get the following correlation matrix,

$$
\mathbf{R}_{3}=\left(\begin{array}{ccc}
1 & 0.5654257 & 0.65043827 \\
0.5654257 & 1 & 0.66211557 \\
0.65043827 & 0.66211557 & 1
\end{array}\right)
$$



Figure 4.6: Total buffer content in light and overload traffic for systems with $K=3$ mobile users with correlated channels (Clayton copula) and $M=3$ states per channel, working under MaxRate(MR), MaxWeight(MW), LCQ, DSP and GPS scheduling disciplines number of expansion terms $N=50$.

Figure 4.6 again compares the MaxRate, MaxWeight and LCQ schedulers as well as the DPS and GPS schedulers. We show the 50th order light-traffic and overload expansions of the total mean queue content, as well as simulation results to verify the accuracy of the approximations. We here use the first correlated channel model with correlation matrix $\mathbf{R}_{1}$. We again obtain a perfect match in a first region, followed by a sharp deviation from the correct value. This again suggests that the region of convergence of the series expansion ends at the position of this fast deviation. Figure 4.7 depicts the same values for the Vine copula with rate matrix $\mathbf{R}_{2}$.

Positive spatial correlations reduce MU diversity and therefore also reduce


Figure 4.7: Total buffer content in light and overload traffic for systems with $K=3$ mobile users with correlated channels (Vine copula) and $M=3$ states per channel, working under MaxRate(MR), MaxWeight(MW), LCQ, DSP and GPS scheduling disciplines; number of expansion terms $N=50$.
the gain of opportunistic scheduling. Nevertheless, if correlations are not strong, the opportunistic approach can still be beneficial. Figure 4.8 demonstrates the diminished efficiency of opportunistic scheduling by comparing the MaxRate, MaxWeight and LCQ schedulers for independent and correlated channels in the overloaded regime. In particular, Figure 4.8 compares the differences in total queue content between MaxRate and LCQ, MaxWeight and LCQ, and GPS and DPS,

$$
\begin{aligned}
\text { MaxRate and LCQ : } & \Delta_{\mathrm{MR}}^{\mathrm{LCQ}}=\mathrm{E}\left[Q^{\mathrm{LCQ}}\right]-\mathrm{E}\left[Q^{\mathrm{MR}}\right], \\
\text { MaxWeight and LCQ : } & \Delta_{\mathrm{MW}}^{\mathrm{LCQ}}=\mathrm{E}\left[Q^{\mathrm{LCQ}}\right]-\mathrm{E}\left[Q^{\mathrm{MR}}\right], \\
\text { GPS and DPS : } & \Delta_{\mathrm{GSP}}^{\mathrm{DSP}}=\mathrm{E}\left[Q^{\mathrm{DPS}}\right]-\mathrm{E}\left[Q^{\mathrm{GPS}}\right] .
\end{aligned}
$$

Note that a positive difference corresponds to a performance gain of the more opportunistic scheduler (MR, MW, DPS). For each comparison, we plot these differences for uncorrelated channels, for the Vine copula channel with rate matrix $\mathbf{R}_{2}$ (weak correlation), and for the Vine copula channel with rate matrix $\mathbf{R}_{3}$ (strong correlation). As previously noted, both MaxWeight and MaxRate outperform the LCQ scheduler, while GPS performs better than DPS, the performance gain being higher for less heavily loaded systems. The performance gain however clearly diminishes by the introduction of channel correlations, more correlation meaning less gain.

### 4.6 Conclusions

We considered a queueing model for assessing the performance of a downlink wireless MU transmission scenario, under varying channel conditions. The buffer behaviour of the wireless access point was modelled by a queueing system with multiple queues and a shared server. Accounting for time-correlation of the channel quality, the channels were modelled by an exogenous Markov process, each state of this Markov process corresponding to fixed (but not necessarily equal) channel qualities of the different channels.

As the state space of this system is very large and does not have additional structure which can be exploited (like product form, G/M/1-type or M/G/1-type), we focused on a numerical approximation approach which relies on series expansion techniques. We showed that this approach can calculate various performance measures fast in the light-traffic and the overload-traffic regimes.

We then adapted a discrete-time Markov model from literature for a single Rayleigh fading channel, first to an equivalent continuous-time Markov channel model, and then to a multi-channel Markov model. In the latter case, we relied on copulas (Clayton and Vine copulas) of the stationary distribution of the SNR of the different channels for the introduction of cross-channel correlation.


Figure 4.8: Difference in total mean buffer content in overload traffic for a system with $K=3$ mobile users under independent, weakly correlated and strongly correlated channels for MaxRate(MR), MaxWeight(MW), LCQ, DPS and GPS scheduling disciplines; the order of the expansion is $N=50$.

For the purpose of demonstration, several well-known scheduling disciplines were analysed numerically and compared with each other. The results for both independent and correlated channels were validated by simulation. It was shown that the approximation of the performance measures was not only computationally efficient, but also very accurate in the light-traffic and the overload-traffic regimes.

## References

[1] J.W. Roberts. A survey on statistical bandwidth sharing. Computer Networks, 45(3):319-332, 2004.
[2] S. Patil, and G. de Veciana. Measurement-based opportunistic scheduling for heterogenous wireless systems. IEEE Transactions on Communications, 57(9):2745-2753, 2009.
[3] X. Liu, E.K.P. Chong, and N.B. Shroff. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. IEEE Journal on Selected Areas in Communications, 19(10):2053-2064, 2001.
[4] R. Knopp and P. Humblet. Information capacity and power control in singlecell multiuser communications. In: Proceedings of IEEE ICC'95 "Gateway to Globalization", volume 1, pages 331-334, June 1995.
[5] X. Liu, E.K.P. Chong, and N.B. Shroff. A framework for opportunistic scheduling in wireless networks. Computer Networks, 41(4):451-474, 2003.
[6] X. Liu, E.K. Chong, and N.B. Shroff. Optimal opportunistic scheduling in wireless networks. In: Proceedings of the 58th IEEE Vehicular Technology Conference, VTC 2003-Fall, volume 3, pages 1417-1421, 2003.
[7] W. Ajib and D. Haccoun. An overview of scheduling algorithms in MIMObased fourth-generation wireless-systems. IEEE Network, 19(5):43-48, 2005.
[8] A. Asadi and V. Mancuso. A survey on opportunistic scheduling in wireless communications. IEEE Communications Surveys \& Tutorials, 15(4):16711688, 2013.
[9] J. Mietzner, R. Schober, L. Lampe, W.H. Gerstacker, and P.A. Hoeher. Multiple-antenna techniques for wireless communications-a comprehensive literature survey. IEEE Communications Surveys \& Tutorials, 11(2):87-105, 2009.
[10] S. Shakkottai, T.S. Rappaport, and P.C. Karlsson. Cross-layer design for wireless networks. IEEE Communications Magazine, 41(10):74-80, 2003.
[11] D. Gesbert, M. Kountouris, R.W. Heath, C.B. Chae, and T. Salzer,. Shifting the MIMO paradigm. IEEE Signal Processing Magazine, 24(5):36-46, 2007.
[12] X. Lin, N.B. Shroff, and R. Srikant. A tutorial on cross-layer optimization in wireless networks. IEEE Journal on Selected areas in Communications, 24(8):1452-1463, 2006.
[13] D. Gesbert and M. Slim-Alouini. How much feedback is multi-user diversity really worth? In: Proceedings of the 2004 IEEE International Conference on Communications, volume 1, pages 234-238, 2004.
[14] P. Chaporkar, A. Proutiere, H. Asnani, and A. Karandikar. Scheduling with limited information in wireless systems. In: Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing, pages 75-84, 2009.
[15] B. Makki and T. Eriksson. Multiuser diversity in correlated Rayleigh-fading channels. EURASIP Journal on Wireless Communications and Networking, 2012(1):1-9, 2012.
[16] M. Chiani, M.Z. Win, and A. Zanella. On the capacity of spatially correlated MIMO Rayleigh-fading channels. IEEE Transactions on Information Theory, 49(10):2363-2371, 2003.
[17] B. Nosrat-Makouei, J. G. Andrews, and R. W. Heath Jr. MIMO interference alignment over correlated channels with imperfect CSI. IEEE Transactions on Signal Processing, 59(6):2783-2794, 2011.
[18] L. Badia, A. Baiocchi, A. Todini, S. Merlin, S. Pupolin, A. Zanella, and M. Zorzi. On the impact of physical layer awareness on scheduling and resource allocation in broadband multicellular IEEE 802.16 systems. IEEE Wireless Communications, 14(1):36-43, 2007.
[19] B. Bellalta, A. Faridi, J. Barcelo, V. Daza, and M. Oliver. Queueing analysis in multiuser multi-packet transmission systems using spatial multiplexing. arXiv preprint, arXiv:1207.3506, 2012.
[20] S. Shakkottai and A. L. Stolyar. Scheduling for multiple flows sharing a time-varying channel: The exponential rule. Translations of the American Mathematical Society-Series 2, 207:185-202, 2002.
[21] W. Ajib and D. Haccoun. An overview of scheduling algorithms in MIMObased fourth-generation wireless systems. IEEE Network, 19(5):43-48, 2005.
[22] B. Sadiq, S.J. Baek, and G. De Veciana. Delay-optimal opportunistic scheduling and approximations: The log rule. IEEE/ACM Transactions on Networking, 19(2):405-418, 2011.
[23] Y. Liu, S. Gruhl, and E.W. Knightly. WCFQ: an opportunistic wireless scheduler with statistical fairness bounds. IEEE Transactions on Wireless Communications, 2(5):1017-1028, 2003.
[24] Z.H. Han and Y.H. Lee. Opportunistic scheduling with partial channel information in OFDMA/FDD systems. In: Proceedings of the 60th IEEE Vehicular Technology Conference, VTC2004-Fall, volume 1, pages 511-514, 2004.
[25] L. Lei, C. Lin, J. Cai, and X. Shen. Performance analysis of wireless opportunistic schedulers using stochastic Petri nets. IEEE Transactions on Wireless Communications, 8(4):2076-2087, 2009.
[26] M.M. Rashid, M.J. Hossain, E.H. Hossain, and V.K. Bhargava. Opportunistic Spectrum Scheduling for Multiuser Cognitive Radio: A Queueing Analysis. IEEE Transactions on Wireless Communications, 8(10):5259-5269, 2009.
[27] H.S. Wang and N. Moayeri. Finite-state Markov channel-a useful model for radio communication channels. IEEE Transactions on Vehicular Technology, 44(1):163-171, 1995.
[28] E. Altman, K.E. Avrachenkov, and R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with applications to queueing models. Advances in Applied Probability, 36(3):839-853, 2004.
[29] K. De Turck, E. De Cuypere, S. Wittevrongel, and D. Fiems. Algorithmic approach to series expansions around transient Markov chains with applications to paired queuing systems. In: Proceedings of the 6th International Conference on Performance Evaluation Methodologies and Tools, pages 3844, 2012.
[30] K. De Turck, E. De Cuypere, and D. Fiems. A Maclaurin-series expansion approach to multiple paired queues. Operations Research Letters, 42(3):203207, 2014.
[31] K. De Turck and D. Fiems. A series expansion approach for finite-capacity processor sharing queues. In: Proc. of the 7th International Conference on Performance Evaluation Methodologies and Tools, pp. 118-125, 2013.
[32] J.B. Lasserre. A Formula for Singular Perturbations of Markov Chains. Journal of Applied Probability, 31(3):829-833, 1994.
[33] W.B. van den Hout. The power-series algorithm. PhD Thesis. University of Tilburg. 1996.
[34] B. Błaszczyszyn, T. Rolski, and V. Schmidt. Advances in Queueing: Theory, Methods and Open Problems, chapter: Light-traffic approximations in queues and related stochastic models. CRC Press, Boca Raton, Florida, 1995.
[35] I. Kovalenko. Rare events in queueing theory. A survey. Queueing systems, 16(1):1-49, 1994.
[36] K. Avrachenkov and J.A. Filar. Analytic Perturbation Theory and Its Applications. SIAM, 2014.
[37] E. Evdokimova, K. De Turck, S. Wittevrongel, and D. Fiems. Efficient Performance Evaluation of Wireless Networks with Varying Channel Conditions. In: Proceedings of the 22nd International Conference on Analytical and Stochastic Modelling Techniques and Applications, pages 59-72, Springer LNCS, 2015.
[38] E. Evdokimova, K. De Turck, S. Wittevrongel, and D. Fiems. An Analytical Performance Evaluation Tool for Wireless Access Points with Opportunistic Scheduling. In: Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools, pages 43-48, 2016.
[39] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. IEEE Transactions on Information Theory, 39(2):466-478, 1993.
[40] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. Scheduling in a queueing system with asynchronously varying service rates. Probability in the Engineering and Informational Sciences, 18(02):191-217, 2004.
[41] K.E. Avrachenkov and M. Haviv Perturbation of null spaces with application to the eigenvalue problem and generalized inverses. Linear Algebra and its Applications, 369:1-25, 2003.
[42] P. Sadeghi, R.A. Kennedy, P. B. Rapajic, and R. Shams. Finite-State Markov Modeling of Fading Channels. IEEE Signal Processing Magazine, 25(5):5780, 2008.
[43] Q. Zhang and S. Kassam, Finite-state Markov model for Rayleigh fading channels. IEEE Transactions on Communications, 47(11):1688-1692, 1999.
[44] K. Zheng, F. Liu, L. Lei, C. Lin, and Y. Jiang. Stochastic performance analysis of a wireless finite-state Markov channel. IEEE Transactions on Wireless Communications, 2(12):1063-1072, 2013.
[45] A. Goldsmith. Wireless communications. Cambridge university press, 2005.
[46] R.B. Nelsen. An introduction to copulas. Springer Science \& Business Media, 2007.
[47] A. Panagiotelis, C. Czado, and H. Joe. Pair copula constructions for multivariate discrete data. Journal of the American Statistical Association, 107(499):1063-1072, 2012.
[48] K. Aas, C. Czadob, A. Frigessic, H. Bakkend. Pair-copula constructions of multiple dependence, Insurance: Mathematics and Economics, 44:182-198, 2009.
[49] H.T. Banks, A. Broido, K. Gayvert, S. Hu, M. Joyner, and K. Link. Simulation and Modeling Related to Computational Science and Robotics Technology, chapter: Simulation Algorithms for Continuous Time Markov Chain Models. IOS Press, 2007.
[50] J. Aráuz, P. Krishnamurthy. A study of different partitioning schemes in first order Markovian models for Rayleigh fading channels. In: Proceedings of the 5th International Symposium on Wireless Personal Multimedia Communications, volume 1, pages 277-281, 2002.

# Internet Provisioning in VANETs: Performance Modelling of Drive-Thru Scenarios 

Ekaterina Evdokimova, Alexey Vinel, Nikita Lyamin and Dieter Fiems<br>submitted to IEEE Transactions on Intelligent Transportation Systems.


#### Abstract

Drive-thru-Internet is a scenario in Cooperative Intelligent Transportation systems (C-ITS) where a road-side unit (RSU) provides multimedia services to vehicles that pass by. The performance of drive-thru-Internet depends on various factors, like data traffic intensity, vehicle traffic density, and radio-link quality within the coverage of RSU, and must be evaluated at the stage of system design in order to fulfil Quality of Service requirements of the customers in C-ITS. In this paper, we present an analytical framework that models downlink traffic in a drive-thru-Internet scenario by means of a multidimensional Markov process: the packet arrivals in the RSU buffer constitute Poisson processes and the transmission times are exponentially distributed. Taking into account the state space explosion problem associated with multidimensional Markov processes, we use iterative perturbation techniques to calculate the stationary distribution of the Markov chain. Our numerical results reveal that the proposed approach yields accurate estimates


of various performance metrics like the mean queue content and the mean packet delay for a wide range of system workloads.

### 5.1 Introduction

The emerging concept of Cooperative Intelligent Transportation Systems (C-ITS) suggests a widespread adoption of information and communication technologies in diverse vehicular applications aimed to increase transport safety, efficiency and comfort. C-ITS vehicles exchange information with each other as well as with roadside infrastructure in a heterogeneous wireless networking environment.

There are number of communication technologies currently under development, that could support connectivity in a vehicular environment [1]. To enable vehicular communications in the Dedicated Short Range Communications (DSRC) 5.9 GHz band, IEEE 802.11 p , which is currently integrated into the recent IEEE 802.11-2012 standard, was introduced by Institute of Electrical and Electronics Engineers (IEEE) [2]. The IEEE 802.11p defines two lower layers of the communication stack: the Physical layer and the Medium Access Control layer. IEEE also introduced WAVE (wireless access in vehicular environment), which defines the overall protocol stack for vehicular communications, including management and security planes [3]. At the same time, in Europe, under the European Commission mandate M/453, European Telecommunications Standards Institute (ETSI) developed a C-ITS protocol stack specified in ETSI EN 302665 [4]. The ETSI C-ITS stack consists of three layers: the access, the networking and transport, and the facilities layer, with a number of management and security protocols specified for all three layers [5]. Apart from DSRC, cellular technologies, like LTE (Long Term Evolution) or the currently being developed 5 G could become a complementing technology choice [1, 6]. The upcoming 5G communications promises both operation in extremely mobile environments (up to $500 \mathrm{~km} / \mathrm{h}$ relative speeds) and highly reliable connectivity with low-latency for vehicle to everything (V2X) scenarios [7].

Infotainment services provided to the drivers and passengers are heavily dependent on the connectivity of the vehicles to the Internet. Broadband cellular networks which provide stable vehicle-to-infrastructure communication links are assumed to be available in urban areas. Rural roads, in contrast, might have only intermittent cellular connectivity, which motivates the consideration of a drivethru scenario where a moving vehicle spends at most a couple of minutes in the coverage area of a roadside unit, an access point or a base station.

In this study we consider downlink communications for data downloading by the vehicles from the RSU. There are variety of ITS applications under current consideration, that require such downlink communications in a drive-thru scenario. Some examples include media downloading, map downloading and updating, and
vehicle software/data provisioning and updating [8]. All of these scenarios, assume infrastructure-to-vehicle (I2V) communications: when in the RSU coverage area, a vehicle should be able to download certain data, e.g. web-page content, map segment update, software update, etc. To provide a certain level of Quality-ofService ( QoS ) to users, models to estimate communication performance in drivethru scenario are required.

In [6], the authors highlight the importance of resource allocation and QoS support in DSRC and LTE vehicular communications. In our study we make an attempt to address this issue by proposing a model that captures the resource allocation process in LTE and DSRC and allows for estimating its performance for a downlink drive-thru scenario. Below we summarise similar efforts done in [9-11] and highlight the main limitations of the state-of-the-art models.

In [10], a TDMA-based (Time-division multiple access) scheduling algorithm for drive-thru scenario was proposed. The main idea of the algorithm is to maximise the total achieved utility by all vehicles under the assumption that each TDMA slot is assigned to only one vehicle. However, the authors make the assumption of the vehicles' constant speed for all the vehicles in the road-side unit (RSU) range. In our study we do not fix the speed of the vehicles assuming the vehicle residence time is exponentially distributed.

In [11] authors present an analytical approach to estimate the mean packet service time, queue length and mean throughput per vehicle. However, the presented model is derived under the assumptions of an infinite capacity buffer at the RSU and fixed length data packets. In [9], an approach on spatially coordinated access to the channel in a drive-thru scenario is presented. Following the approach that the coverage area is divided in zones by achievable throughput, the authors propose an algorithm that optimises the assignment of zones to vehicles such that overall system throughput is maximised. These authors demonstrate that spatial optimisation outperforms standard IEEE 802.11 CSMA/CA (Carrier-sense multiple access with collision avoidance) in terms of throughput. In contrast to the aforementioned models, the present model allows us to additionally retrieve the influence of the limitations implied by the limited RSU buffer capacity and to estimate the probabilities that a vehicle will not be served or that a packet will be rejected. In addition, our approach is independent from the underlying communication technology.

To assess the achievable QoS level in drive-thru scenario for different applications, the models should capture various scenario attributes and practical limitations. Since realistic modelling of the drive-thru scenario requires to consider diverse vehicle density and changing speeds, varying channel quality, and realistic RSU configuration (buffer capacities, number of queues, etc.), the problem becomes multi-dimensional. Simulation experiments could become quite extensive, which makes it highly relevant to design an analytical model that could capture
the realistic system behaviour and give a detailed estimation of the communication performance of the system. To draw conclusions on the QoS levels required to support different applications, apart from expected throughput, it could also be necessary to retrieve the probability that a vehicle will not be serviced by the RSU due to its limited resources (packet losses due to the RSU buffer overflow) or due to the fact that the vehicle leaves the RSU coverage zone prior to delivery as well as estimates of queueing specific performance measures like the packet delays and the RSU queue size.

To study scheduling of packets for the drive-thru scenario, we propose a Markov chain model with a multidimensional state space and rely on numerical evaluation techniques for calculating the invariant distribution of the chain. The size of the state space of a multi-dimensional Markov chain being equal to the product of the sizes of the individual dimensions, multi-dimensionality often leads to the so called state-space explosion problem, sometimes also referred to as the curse of dimensionality. As a consequence, already for chains with but a few dimensions, a direct numerical solution of the balance equations is not computationally feasible. For the Markov chain at hand, we combine series expansion techniques and iterative solution methods for matrix equations to find various performance measures like the mean packet delay and the mean buffer content at the RSU. Series expansion techniques for Markov chains are referred to as perturbation techniques, the power series algorithm and light-traffic approximations. While the naming is not absolute, perturbation methods mainly relate to sensitivity analysis of performance measures with respect to various system parameters. In particular singular perturbations have received considerable attention in literature, see [14-16] and the references therein. For such perturbations, the class-structure of the non-perturbed chain is not retained, which poses mathematical complications. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by an auxiliary variable $\rho$. For $\rho=0$, the chain cannot only be solved efficiently, but one can also calculate the series expansion of the solution of the chain in $\rho$. For $\rho=1$ the original Markov chain is retrieved such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of this expansion includes $\rho=1$, see e.g. [17-20]. Finally, light-traffic approximations often corresponds to a series expansion in the arrival rate at a queue. For an overview on series expansion techniques in stochastic systems, see the surveys in [21] and [22]. In the present study we combine expansion techniques and iterative methods, a combination which was previously explored in [23] in the context of stochastic modelling of assembly systems. In addition, we study the model in overload. Overload analysis, the opposite of light-traffic analysis, studies the performance in the limit of the service times growing to infinity, see e.g. [24, 25].

The contributions of this paper could be summarised as follows.

1. We introduce a Markovian queueing model of an infrastructure vehicular network where an RSU transmits data to vehicles in its coverage area which consists of zones with different channel conditions under realistic assumptions of limited buffer length and limited number of customers in service.
2. We tackle the model numerically despite the large-scale of the queueing system at hand. The solution relies on the Taylor series approximation technique and benefits from the sparsity of the transition rate matrix.
3. We conduct performance evaluation of a drive-thru-Internet scenario and quantitatively characterise the mean queue content, the mean packet delay and two types of packet loss: discarded packets and rejected packets.

The remainder of this paper is organised as follows. The next section introduces the Markov process model for studying the drive-thru scenario. Numerical solution techniques for this process are presented in section 5.3 Finally, we evaluate various drive-thru scenarios in section 5.4 and conclude in section 5.5,

### 5.2 Markov model

We propose a Markov model for the drive-through scenario. Following the models presented in [9, 10], we assume drive-thru scenario. A common modelling approach to capture the receiving signal strength increase as long as the vehicle approaches the roadside transmitter in the drive-thru scenario is to split the cell into the production zone where the network throughput is high as well as the entry and exit zones [9, 10] where the channel quality is poor, Figure [5.1] Such a simple assumption enables the analysis independently of the underlying communication technology e.g. DSRC or LTE. Multiple cars move from left to right through different zones. We focus on RSU to vehicle station communication. In this work we do not consider the service discovery process and assume it established successfully for each vehicle when entering the zone (for service discovery studies interested users may refer to [12, 13]). A single RSU communicates with all the vehicles in the different zones, the channel conditions differing from zone to zone. We make the following assumptions.

### 5.2.1 Modelling assumptions

Let $\mathcal{M}=\{1, \ldots, M\}$ be the set of zones, $M>0$ denoting the number of zones. Vehicles move from zone 1 to zone $M$, thereby traversing all the zones in numerical order. The number of vehicles in zone $m \in \mathcal{M}$ at time $t$ is denoted by $V_{m}(t)$. Following the model presented in [10] we assume a new vehicle arrives at zone 1 in accordance with a Poisson process with rate $\alpha$. Each vehicle remains for an


Figure 5.1: Drive-thru scenario. Multiple cars move from left to right through different zones, each zone representing different channel conditions.
exponentially distributed time in each zone. The mean residence time in zone $m$ is denoted by $1 / \beta_{m}$ and the vehicle moves to zone $m+1$ upon departing zone $m$, for $m \in \mathcal{M} \backslash\{M\}$. When the vehicle leaves zone $M$ it leaves the range of the RSU. We further impose an upper bound $K$ on the number of cars that can be simultaneously served by the RSU,

$$
\sum_{m=1}^{M} V_{m}(t) \leq K
$$

New vehicles arriving in zone 1 will not be able to connect to the RSU when there are already $K$ vehicles connected to the RSU. These units will also not connect if other vehicles leave the transmission range of the RSU while they are still in range.

For each vehicle in range, packets arrive in accordance with a Poisson process with rate $\lambda$ and are stored in a dedicated buffer at the RSU with finite capacity $C$. Let $X_{(n, m)}(t)$ denote the number of vehicles in zone $m$ that have $n$ packets in the buffer at the RSU at time $t$, for $(n, m) \in \mathcal{C}$,

$$
\mathcal{C}=\{0,1, \ldots, C\} \times \mathcal{M}
$$

Furthermore, let $\mathbf{X}(t)=\left[X_{(n, m)}(t)\right]_{(n, m) \in \mathcal{C}}$. Note that $V_{m}(t)$ can be expressed in terms of $X_{(n, m)}(t)$,

$$
V_{m}(t)=\sum_{n=0}^{C} X_{(n, m)}(t),
$$

as $V_{m}(t)$ includes all vehicles in zone $m$, regardless of the number of waiting packets at the RSU.

The packet length is exponentially distributed. However, as the communication channel is shared by all vehicles in range, the rate at which packets leave
the RSU buffers depends on the complete system state: the number of vehicles in the different zones, and the number of outstanding packets for each vehicle. For $\mathbf{X}(t)=\mathbf{x}=\left[x_{(n, m)}\right]_{(n, m) \in \mathcal{C}}$, a packet is transmitted successfully in the time interval $(t, t+d t]$ to a particular vehicle in zone $m$ with $n$ packets waiting at the RSU with probability $\mu_{(n, m)}(\mathbf{x}) d t+o(d t)$. In other words, there is a packet transmitted to an vehicle in zone $m$ with $n$ packets waiting at the RSU with rate $x_{(n, m)} \mu_{(n, m)}(\mathbf{x})$ when the system is in state $\mathbf{x}$. Of course there are no packet departures to vehicles in zone $m$ with $n$ packets in the buffer, if there are no such vehicles $\left(x_{(n, m)}=0\right)$.

In view of the assumptions on the vehicle arrivals and residence times in the different zones, and in view of the assumptions on packet arrivals and departures, the process $\{\mathbf{X}(t), t \in \mathbb{R}\}$ constitutes a Markov process. The state space of this process is,

$$
\mathcal{K}=\left\{\mathbf{x} \in \mathbb{N}^{(C+1) M},|\mathbf{x}|_{1} \leq K\right\}
$$

where $|x|_{1}=\sum_{(n, m) \in \mathcal{C}} x_{(n, m)}$ denotes the Manhattan norm of the vector $\mathbf{x}$.

### 5.2.2 Balance equations

We now focus on the balance equations of the Markov chain $\{\mathbf{X}(t), t \in \mathbb{R}\}$. For ease of notation, let $\mathbf{e}_{(n, m)}$ be a vector with all elements equal to zero apart from the element with index $(n, m)$ which is one. We have the following possible state transitions.

- Arrivals of new vehicles: for any state $\mathbf{x} \in \mathcal{K}$ such that $|\mathbf{x}|_{1}<K$, new vehicles arrive in zone 1 with constant rate $\alpha$. The new state is $\mathbf{x}+\mathbf{e}_{(0,1)}$. That is, the number of vehicles with an empty buffer in zone 1 increases by 1. For states $\mathbf{x}$ with $|\mathbf{x}|_{1}=K$, no new vehicles can connect with the RSU, hence there are no new arrivals of vehicles .
- Arrivals of new packets: for any state $\mathbf{x} \in \mathcal{K}$ and any $(n, m) \in\{0, \ldots, C-$ $1\} \times \mathcal{M}$, the total arrival rate for all vehicles with $n$ outstanding packets in zone $m$ is $x_{(n, m)} \lambda$ as there are $x_{(n, m)}$ vehicles, each receiving packets with constant rate $\lambda$. The arrival invokes a state transition to state $\mathbf{x}+\mathbf{e}_{(n+1, m)}-$ $\mathbf{e}_{(n, m)}$. Note that the first index only runs till $C-1$ as the buffer of the vehicles with $C$ outstanding packets is full and new packet arrivals are not accepted.
- vehicles moving zones: for any state $\mathbf{x}$ and for any $(n, m) \in \mathcal{C}$, an vehicle with $n$ packets moves from zone $m$ to zone $m+1$ with rate $x_{(n, m)} \beta_{m}$ (as there are $x_{(n, m)}$ such vehicles, each of these moving to zone $m+1$ with rate $\beta_{m}$ ). This invokes a state transition to state $\mathbf{x}+\mathbf{e}_{(n, m+1)}-\mathbf{e}_{(n, m)}$ for $m<M$ and a state transition to $\mathbf{x}-\mathbf{e}_{(n, m)}$ for $m=M$. In the latter case, the vehicle leaves the transmission range of the RSU.
- Departure of a packet: for a given $\mathbf{x}$, there is a packet transmisison to an vehicle with $n$ outstanding packets in zone $m$ with rate $\mu_{(n, m)}(\mathbf{x})$. Hence the total transmission rate to all vehicles $n$ outstanding packets in zone $m$ with is $x_{(n, m)} \mu_{(n, m)}(\mathbf{x})$. When there is such a transmission, the new state is $\mathbf{x}+\mathbf{e}_{(n-1, m)}-\mathbf{e}_{(n, m)}$ (one vehicle with $n$ outstanding packets less, and one vehicle with $n-1$ outstanding packets more in zone $m$ ).

Let $\pi(\mathbf{x})=\lim _{t \rightarrow \infty} \mathrm{P}[\mathbf{X}(t)=\mathbf{x}]$ be the stationary probability to be in state $\mathrm{x} \in \mathcal{K}$. In view of the possible transitions discussed above, the balance equations read,

$$
\begin{align*}
& \pi(\mathbf{x}) \gamma_{1}(\mathbf{x})=\alpha \pi\left(\mathbf{x}-\mathbf{e}_{(0,1)}\right) \\
&+\lambda \sum_{n=1}^{C} \sum_{m=1}^{M}\left(x_{\left(n^{-}, m\right)}+1\right) \pi\left(\mathbf{x}+\mathbf{e}_{\left(n^{-}, m\right)}-\mathbf{e}_{(n, m)}\right) \\
&+ \sum_{n=0}^{C} \sum_{m=2}^{M} \beta_{m^{-}}\left(x_{\left(n, m^{-}\right)}+1\right) \pi\left(\mathbf{x}+\mathbf{e}_{\left(n, m^{-}\right)}-\mathbf{e}_{(n, m)}\right) \\
&+\beta_{M} \sum_{n=0}^{C}\left(x_{(n, M)}+1\right) \pi\left(\mathbf{x}+\mathbf{e}_{(n, M)}\right) \\
&+ \sum_{n=0}^{C-1} \sum_{m=1}^{M}\left(x_{\left(n^{+}, m\right)}+1\right) \mu_{n, m}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}\right) \\
& \times \pi\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}-\mathbf{e}_{(n, m)}\right) \tag{5.1}
\end{align*}
$$

for $\mathbf{x} \in \mathcal{K}$, with $n^{+}=n+1, n^{-}=n-1, m^{+}=m+1$, and $m^{-}=m-1$ to simplify the notation and with,

$$
\begin{aligned}
& \gamma_{1}(\mathbf{x})=\alpha 1_{\left\{|\mathbf{x}|_{1}<K\right\}} \\
&+\sum_{n=0}^{C} \sum_{m=1}^{M} x_{(n, m)}\left(\lambda 1_{\{n<C\}}+\mu_{(n, m)}(\mathbf{x})+\beta_{m}\right) .
\end{aligned}
$$

The set of balance equations allows for determining all unknown probabilities $\pi(\mathbf{x}), \mathbf{x} \in \mathcal{K}$ up to a factor. The remaining unknown factor follows from the normalisation condition,

$$
\sum_{\mathbf{x} \in \mathcal{K}} \pi(\mathbf{x})=1
$$

The solution of the set of equations is unique as it is easily verified that the Markov process is ergodic provided that $\alpha>0, \lambda>0$ and $\beta_{m}>0$ for $m \in \mathcal{M}$.

### 5.2.3 Discussion

Each vehicle within range can be in $M(C+1)$ different states: the vehicle is in 1 out of $M$ possible zones with 0 up to $C$ waiting packets at the RSU. There are


Figure 5.2: State space size versus maximum number of customers $K$ in the system with $M=3$ zones and buffer capacity $C=1,3$ and 5.
$K$ vehicles, which are either in range or not, so each vehicle can be in $M(C+$ 1) +1 states, the additional state corresponding to the case the vehicle is not in range. Therefore, the total number of possible states is equal to the number of $K$-combinations with repetition [26] (as we only track the number of vehicles in the different states, and not the state of each MU),

$$
\begin{align*}
N_{s} & =\left(\binom{M(C+1)+1}{K}\right)=\binom{K+M(C+1)}{K} \\
& =\frac{(K+1)(K+2) \cdots(K+M(C+1))}{(M(C+1))!} \tag{5.2}
\end{align*}
$$

To illustrate the impact of $C$ and $K$ on the size of the state space, figure 5.2 shows the size of the state space vs. $K$ for different values of $C$ as indicated, and for $M=3$ zones. The figure 5.2 reveals that the state space quickly grows with the parameters $K$ and $C$. Only for very small $K$ and $C$ (and $M$ ), it is computationally feasible to solve the system of balance equations directly.

While a direct solution is not computationally feasible, the number of possible transitions from any state is at most $3 K$ (for each MU, there can be an arrival, a departure or a change of zone) which is far smaller than the size of the state space. For such sparse systems of equations, iterative solution methods like Jacobi iteration, the Gauss-Seidel method or successive overrelaxation are effective, the complexity of a single iteration being $O\left(N_{s} K\right)$. In the remainder, we rely on these methods as well as on power-series expansion techniques. Series expansion techniques are used to obtain performance measure in a parameter range, rather than in a single point in the parameter space. Moreover, we identify a number of cases in which series expansion techniques can be applied with the same numerical complexity of the iterative solution methods.

With the above mentioned methods, the effects of the state space explosion problem on the solution speed are mitigated, but $M, K$ and $C$ still cannot be chosen arbitrarily large. This is the case for any numerical solution method as such methods are at least $O\left(N_{s}\right)$ as one needs to calculate $N_{s}$ stationary probabilities. While the number of zones $M$ and the number of vehicles $K$ are typically limited, the buffer capacity $C$ typically is fairly large. The state-space explosion problem being very present for large $C$, the following interpretation of the buffer size is introduced to approximate systems with larger buffers.

Rather than precisely tracking the number of packets in the buffer for each MU, we only approximately track the queue content, each queue content level in the model representing a range of the queue content in reality. Equivalently, we can interpret this as a merger of multiple packets into a single superpacket, which obviously now requires more time to transmit. Let $C_{\mathrm{MB}}$ denote the size of the buffer in Megabyte. We then set the mean superpacket size to $C_{\mathrm{MB}} / C$, a larger $C$ corresponding to a more precise model (which is also harder to solve). The transmission time of the superpacket being linear in the size of the packet, the transmission rate of the superpacket equals,

$$
\widehat{\mu}_{(n, m)}(\mathbf{x})=\mu_{(n, m)}(\mathbf{x}) \frac{C_{\mathrm{MB}}}{C \mathbb{E}\left[P_{\mathrm{MB}}\right]}
$$

where $\mathbb{E}\left[P_{\mathrm{MB}}\right]$ denotes the mean (original) packet size in Megabyte and where $\mu_{(n, m)}(\mathbf{x})$ denotes the rate corresponding to this packet size.

As the superpacket model is essentially the same as the original model, we will not explicitly introduce superpackets in the remainder of the analysis. The queue capacity $C$ in the numerical examples will be limited though, and the packet size will be larger than the typical IP packet. In other words, for the numerical examples, we will indeed merge multiple IP packets to facilitate the performance analysis.

### 5.3 Performance analysis

Before introducing the various series approximations, we introduce some additional notation. For scaling how vehicles move from zone to zone and for scaling the transmission times, we express the rates to move zones and the transmission rates as follows,

$$
\beta_{m}=\beta \widehat{\beta}_{m}, \quad \mu_{(n, m)}(\mathbf{x})=\mu \widehat{\mu}_{(n, m)}(\mathbf{x}),
$$

where $\widehat{\beta}_{m}$ and $\widehat{\mu}_{(n, m)}(\mathbf{x})$ denote the moving and transmission rates before scaling. We assume that these unscaled rates are given and now study the effects of $\alpha, \beta, \lambda$ and $\mu$ on the system dynamics.

### 5.3.1 A few vehicles with high load

We first investigate performance when the arrival rate of new vehicles is low ( $\alpha$ close to zero) and the transmission rate of the vehicles is low ( $\mu$ close to zero). To this end, we set $\alpha \doteq \mu$ and calculate the series expansion of the stationary distribution vector $\boldsymbol{\pi}$ for $\mu \rightarrow 0$.

Plugging $\mu_{(n, m)}(\mathbf{x})=\mu \widehat{\mu}_{(n, m)}(\mathbf{x})$ and $\alpha=\mu$ and the series expansion,

$$
\pi(\mathbf{x})=\sum_{i=0}^{\infty} \pi_{i}(\mathbf{x}) \mu^{i}
$$

into the balance equations (5.1) and isolating the terms in $\mu^{i}$ leads to,

$$
\begin{align*}
\pi_{i}(\mathbf{x})= & \frac{1}{\gamma_{2}(\mathbf{x})}\left(-\pi_{i-1}(\mathbf{x}) 1_{\left\{|\mathbf{x}|_{1}<K\right\}}\right. \\
& \left.-\pi_{i-1}(\mathbf{x}) \sum_{n=0}^{C} \sum_{m=1}^{M} x_{(n, m)} \widehat{\mu}_{(n, m)}(\mathbf{x})\right)+\pi_{i-1}\left(\mathbf{x}-\mathbf{e}_{(0,1)}\right) \\
& +\lambda \sum_{n=1}^{C} \sum_{m=1}^{M}\left(x_{\left(n^{-}, m\right)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n^{-}, m\right)}-\mathbf{e}_{(n, m)}\right) \\
& +\sum_{n=0}^{C} \sum_{m=2}^{M} \beta_{m-1}\left(x_{\left(n, m^{-}\right)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n, m^{-}\right)}-\mathbf{e}_{(n, m)}\right) \\
& +\beta_{M} \sum_{n=0}^{C}\left(x_{(n, M)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{(n, M)}\right) \\
& +\sum_{n=0}^{C-1} \sum_{m=1}^{M}\left(x_{\left(n^{+}, m\right)}+1\right) \widehat{\mu}_{n, m}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}\right) \\
& \left.\times \pi_{i-1}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}-\mathbf{e}_{(n, m)}\right)\right) \tag{5.3}
\end{align*}
$$

for $\mathbf{x} \in \mathcal{K}$ and $i \in \mathbb{N}^{+}$, and with

$$
\gamma_{2}(\mathbf{x})=\sum_{n=0}^{C} \sum_{m=1}^{M} x_{(n, m)}\left(\lambda 1_{\{n<C\}}+\beta_{m}\right)
$$

For $i=0$, it is easily verified that,

$$
\pi_{0}(\mathbf{x})= \begin{cases}1 & \text { for } \mathbf{x}=\mathbf{0} \\ 0 & \text { otherwise }\end{cases}
$$

Indeed, for $\alpha=0$ there are no arrivals of new vehicles. As $\beta_{m}>0$, all vehicles in range eventually leave, such that there are no vehicles in stationary regime.

The system of equations (5.3), along with the normalisation condition,

$$
\sum_{\mathbf{x} \in \mathcal{K}} \pi_{i}(\mathbf{x})=1_{\{i=0\}}
$$

allows for solving for all $\pi_{i}(\mathbf{x})$ once the terms of order $(i-1)$ are determined. Moreover, solving this system of equations is computationally less demanding, compared to the original balance equations (5.1).

To show this, we first define the following order for the state space $\mathcal{K}$. The dimensions of the state space have a double index. Let the index $(n, m)$ correspond to position $n M+m$ in the state vector which corresponds to the lexicographical order of the pairs $(n, m)$. Given this order, we order the state vectors in lexicographical order as well. Close inspection of equation (5.3) now shows that to calculate $\pi_{i}(\mathbf{x})$, we only need the $i$ th order terms for states which are larger than $\mathbf{x}$. This means that we can calculate the terms $\pi_{i}(\mathbf{x})$ one by one in reverse order. This is possible for all states apart from state $\mathbf{x}=\mathbf{0}$ as $\gamma_{2}(\mathbf{0})=0$. In this case, we invoke the normalisation condition, which leads to,

$$
\pi_{i}(\mathbf{0})=-\sum_{\mathbf{x} \in \mathcal{K} \backslash\{\mathbf{0}\}} \pi_{i}(\mathbf{x})
$$

The numerical complexity for the calculation of single term in the expansion is $O\left(N_{s} K\right)$, which corresponds to the complexity of calculating a single iteration using the Gauss-Seidel method. Whereas every iteration in the Gauss-Seidel method improves the accuracy of the solution in a single point in parameter space, each "iteration" of the perturbation improves the accuracy of the solution in a region in the parameter space around the point $\mu=0$.
Remark 10. The existence of a series expansion in a region around $\mu=0$ is guaranteed as we only have a finite number of equations, and the solution for $\mu=0$ is well defined. Indeed, by Cramer's rule, we know that the stationary distribution is a rational function of $\mu$. Such functions only have a finite number of isolated singularities, $\mu=0$ not being one as there is a unique solution for $\mu=0$. Hence, the stationary probabilities are analytic functions of $\mu$ in a region around $\mu=0$.

### 5.3.2 Perturbation of the arrival rate: $\lambda \approx \lambda_{0}$

We study the stationary distribution $\pi$ as a function of $\lambda$ in a neighbourhood of a fixed arrival rate $\lambda_{0}$, while keeping all other rates constant. To this end, consider the following Taylor series expansion of the stationary probabilities,

$$
\begin{equation*}
\pi(\mathbf{x})=\sum_{i=0}^{\infty} \pi_{i}(\mathbf{x})\left(\lambda-\lambda_{0}\right)^{i}, \tag{5.4}
\end{equation*}
$$

for $\mathrm{x} \in \mathcal{K}$, and $\lambda_{0} \geq 0$. As in the preceding subsection, one easily shows the existence and convergence of the series expansion in a region around $\lambda_{0}$ by Cramer's rule and the uniqueness of the solution for $\lambda=\lambda_{0}$.

Plugging the series expansion (5.4) into the balance equations (5.1) and isolating the terms in $\left(\lambda-\lambda_{0}\right)^{i}$ yields,

$$
\begin{equation*}
\pi_{i}(\mathbf{x})=\frac{\phi_{i}(\mathbf{x})}{\gamma_{3}(\mathbf{x})} \tag{5.5}
\end{equation*}
$$

for $\mathbf{x} \in \mathcal{K}$ and $i \in \mathbb{N}^{+}$, and with,

$$
\begin{aligned}
& \phi_{i}(\mathbf{x})=-\pi_{i-1}(\mathbf{x}) \sum_{n=0}^{C-1} \sum_{m=1}^{M} x_{(n, m)}+\alpha \pi_{i}\left(\mathbf{x}-\mathbf{e}_{(0,1)}\right) \\
&+\sum_{n=1}^{C} \sum_{m=1}^{M}\left(x_{\left(n^{-}, m\right)}+1\right) \pi_{i-1}\left(\mathbf{x}+\mathbf{e}_{\left(n^{-}, m\right)}-\mathbf{e}_{(n, m)}\right) \\
&+\lambda_{0} \sum_{n=1}^{C} \sum_{m=1}^{M}\left(x_{\left(n^{-}, m\right)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n^{-}, m\right)}-\mathbf{e}_{(n, m)}\right) \\
&+\sum_{n=0}^{C} \sum_{m=2}^{M}\left(x_{\left(n, m^{-}\right)}+1\right) \beta_{m^{-}} \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n, m^{-}\right)}-\mathbf{e}_{(n, m)}\right) \\
&+\beta_{M} \sum_{n=0}^{C}\left(x_{(n, M)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{(n, M)}\right) \\
&+\sum_{n=0}^{C-1} \sum_{m=1}^{M}\left(x_{\left(n^{+}, m\right)}+1\right) \mu_{(m, n)}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}\right) \\
& \quad \times \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}-\mathbf{e}_{(n, m)}\right)
\end{aligned}
$$

and,

$$
\gamma_{3}(\mathbf{x})=\alpha 1_{\left\{|\mathbf{x}|_{1}<K\right\}}+\sum_{n=0}^{C} \sum_{m=1}^{M} x_{(n, m)}\left(\lambda_{0} 1_{\{n<C\}}+\beta_{m}+\mu_{(m, n)}(\mathbf{x})\right)
$$

In contrast to the preceding section, the system of equations (5.5) cannot be easily solved. Therefore, we rely on the successive overrelaxation (SOR) method [27]. SOR is an iterative method, which updates the values $\pi_{i}(\mathbf{x})$ in accordance with,

$$
\begin{equation*}
\pi_{i}(\mathbf{x}) \leftarrow(1-\omega) \pi_{i}(\mathbf{x})+\omega \frac{\phi_{i}(\mathbf{x})}{\gamma_{3}(\mathbf{x})} \tag{5.6}
\end{equation*}
$$

for all $\mathbf{x} \in \mathcal{K}$. Here $\omega>0$ is the relaxation factor, a sufficiently small value allowing one to ensure convergence of the method, while a larger value can speed up convergence. Note that $\phi_{i}(\mathbf{x})$ implicitly depends on the values $\pi_{i}(\mathbf{y}), \mathbf{y} \in \mathcal{K}$, the most recently calculated value of $\pi_{i}(\mathbf{y})$ being used in the calculations above.

Summarising, we first calculate $\pi_{0}(\mathbf{x})$. To this end, we initialise $\pi_{0}(\mathbf{0})=$ $1_{\{\mathbf{x}=\mathbf{0}\}}$ and then update $\pi_{0}(\mathbf{x})$ by 5.6 with $i=0$ for all $\mathbf{x} \in \mathcal{K}$, using the order of the state space introduced in subsection 5.3.1 Note that (5.6) also holds for $i=0$ if we set $\pi_{-1}(\mathbf{x})=0$ for all $\mathbf{x} \in \mathcal{K}$, see (5.1). We then repeat updating the values $\pi_{0}(\mathbf{x})$ till convergence.

For the higher order terms, we can assume that $\pi_{i-1}(\mathbf{x}), \mathbf{x} \in \mathcal{K}$ is already known. We then proceed as for the 0 th order term. Set $\pi_{0}(\mathbf{0})=0$ for all $\mathbf{x} \in \mathcal{K}$, and repeat updating these values by (5.6) till we have convergence.

### 5.3.3 Perturbation of the transmission rate $\mu \approx \mu_{0}$

We can retrieve a similar perturbation in the transmission rate $\mu$. To this end, we consider the series expansion,

$$
\pi(\mathbf{x})=\sum_{\mathbf{x} \in \mathcal{K}} \pi_{i}(\mathbf{x})\left(\mu-\mu_{0}\right)^{i}
$$

We follow the same approach as in the preceding section, again relying on the SOR to iteratively solve the set of equations for the terms in the expansion. As for the $\lambda$-perturbation, we choose the initial values for the $i$ th order term as follows,

$$
\pi_{i}(\mathbf{x})=1_{\{\mathbf{x}=\mathbf{0}, i=0\}}
$$

We then update the values $\pi_{i}(\mathbf{x})$, in accordance with,

$$
\begin{aligned}
& \pi_{i}(\mathbf{x}) \leftarrow(1-\omega) \pi_{i}(\mathbf{x})+\omega \frac{1}{\gamma_{4}(\mathbf{x})}\left(\alpha \pi_{i}\left(\mathbf{x}-\mathbf{e}_{(0,1)}\right)\right. \\
& \quad-\pi_{i-1}(\mathbf{x}) \sum_{n=1}^{C} \sum_{m=1}^{M} x_{(n, m)} \widehat{\mu}_{(m, n)}(\mathbf{x}) \\
& \quad+\lambda \sum_{n=1}^{C} \sum_{m=1}^{M}\left(x_{\left(n^{-}, m\right)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n^{-}, m\right)}-\mathbf{e}_{(n, m)}\right) \\
& \quad+\sum_{n=0}^{C} \sum_{m=2}^{M}\left(x_{\left(n, m^{-}\right)}+1\right) \beta_{m^{-}} \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n, m^{-}\right)}-\mathbf{e}_{(n, m)}\right) \\
& \quad+\beta_{M} \sum_{n=0}^{C}\left(x_{(n, M)}+1\right) \pi_{i}\left(\mathbf{x}+\mathbf{e}_{(n, M)}\right) \\
& \quad+\sum_{n=0}^{C-1} \sum_{m=1}^{M}\left(x_{\left(n^{+}, m\right)}+1\right) \widehat{\mu}_{(m, n)}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}\right) \\
& \quad \times \pi_{i-1}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}-\mathbf{e}_{(n, m)}\right) \\
& \quad+\sum_{n=0}^{C-1} \sum_{m=1}^{M}\left(x_{\left(n^{+}, m\right)}+1\right) \mu_{0} \widehat{\mu}_{(m, n)}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}\right)
\end{aligned}
$$

$$
\left.\times \pi_{i}\left(\mathbf{x}+\mathbf{e}_{\left(n^{+}, m\right)}-\mathbf{e}_{(n, m)}\right)\right)
$$

with

$$
\gamma_{4}(\mathbf{x})=\alpha 1_{\left\{|\mathbf{x}|_{1}<K\right\}}+\sum_{n=0}^{C} \sum_{m=1}^{M} x_{(n, m)}\left(\lambda 1_{\{n<C\}}+\beta_{m}+\mu_{0} \widehat{\mu}_{(m, n)}(\mathbf{x})\right)
$$

for $\mathbf{x} \in \mathcal{K}$ and $i \in \mathbb{N}$. Here we define $\pi_{-1}(\mathbf{x})=0$ for all $\mathbf{x} \in \mathcal{K}$. The value $\omega>0$ is again the relaxation factor of the SOR method, a sufficiently small value allowing one to ensure convergence of the method, while a larger value can speed up convergence.

### 5.3.4 Performance metrics

The former calculations allow for approximating the stationary distribution $\pi$ in a region of the parameter space. We now express various performance measures in terms of this stationary distribution. Some relevant performance measures can also be expressed in terms of the stationary distribution of a Markov chain with a smaller state space. This Markov chain is also introduced below.

### 5.3.4.1 Moments of the number of vehicles, and the vehicle blocking probability

Vehicles move through the system, independently from the number of outstanding packets at the RSU. Hence, any performance metric relating to the number of vehicles in the system, can be expressed in terms of the stationary distribution of a Markov chain that only tracks the position of the vehicles, but not the number of packets for these vehicles.

Let $V_{m}(t)=\sum_{n=0}^{C} X_{(n, m)}(t)$ again denote the number of vehicles in zone $m$ at time $t$, and let $\mathbf{V}(t)=\left[V_{m}(t)\right]_{m}=\mathbf{1}^{M}$ be the vector with entries $V_{m}(t)$. The process $\mathbf{V}(t)$ is a Markov process as the evolution of the position of the vehicles does not depend on the queue content of these vehicles. For $\mathbf{y}=\left[y_{1}, \ldots, y_{M}\right] \in$ $\left\{\mathbf{z} \in \mathbb{N}^{M} ;|\mathbf{z}|_{1} \leq K\right\} \doteq \mathcal{N}$, let $\widehat{\pi}(\mathbf{y})=\lim _{t \rightarrow \infty} \mathrm{P}[\mathbf{V}(t)=\mathbf{y}]$ denote the stationary distribution of this Markov chain. We easily obtain the following set of balance equations,

$$
\begin{align*}
& \widehat{\pi}(\mathbf{y})\left(\alpha 1_{\left\{|\mathbf{y}|_{1}<K\right\}}+\sum_{m=1}^{M} y_{m} \beta_{m}\right)=\alpha \widehat{\pi}\left(\mathbf{y}-\mathbf{e}_{1}\right) \\
& +\sum_{m=1}^{M-1} \beta_{m}\left(y_{m}+1\right) \widehat{\pi}\left(\mathbf{y}+\mathbf{e}_{m}-\mathbf{e}_{m+1}\right) \\
&  \tag{5.7}\\
& \quad+\beta_{M}\left(y_{M}+1\right) \widehat{\pi}\left(\mathbf{y}+\mathbf{e}_{M}\right)
\end{align*}
$$

for $\mathbf{y} \in \mathcal{N}$, with normalisation condition,

$$
\sum_{\mathbf{y} \in \mathcal{N}} \widehat{\pi}(\mathbf{y})=1
$$

Here, $\mathbf{e}_{m}$ denotes a vector of all zeroes, except the $m$ th element which is equal to 1 .
The state space of the Markov chain $\mathbf{M}(t)$ is far smaller than the state space of the Markov chain $\mathbf{Q}(t)$. Using similar arguments as in section5.2.3, we find that the size of the state space equals,

$$
N_{s}=\left(\binom{M+1}{K}\right)=\binom{K+M}{K}
$$

As the size is limited, the system of equations 5.7) is easily solved.
Let $V$ denote the number of vehicles connected to the RSU. Once the solution is found, the mean number of vehicles $\mathbb{E}[V]$ that are connected to the RSU, can be expressed in terms of the stationary probabilities $\widehat{\pi}(\mathbf{y})$ as,

$$
\mathbb{E}[V]=\sum_{\mathbf{y} \in \mathcal{N}} \widehat{\pi}(\mathbf{y})|\mathbf{y}|_{1} .
$$

As vehicles cannot connect if already $K$ vehicles are connected, we have that the blocking probability $P_{b}$ equals the probability that there are $K$ vehicles connected $\mathrm{P}[G=K]$, or,

$$
P_{b}=\mathrm{P}[V=K]=\sum_{\mathbf{y} \in \mathcal{N}} \widehat{\pi}(\mathbf{y}) 1_{\{|\mathbf{y}|=K\}} .
$$

### 5.3.4.2 Mean queue size at the RSU

Let $Q_{\text {total }}$ denote the number of packets waiting at the RSU for all vehicles. Its mean can then be expressed in terms of the stationary distribution $\pi$ as,

$$
\mathbb{E}\left[Q_{\text {total }}\right]=\sum_{\mathbf{x} \in \mathcal{K}} \sum_{n=0}^{C} \sum_{m=1}^{M} \pi(\mathbf{x}) x_{(n, m)} n .
$$

The mean number of packets at the RSU for a single vehicle $\mathbb{E}[Q]$ then relates to $\mathbb{E}\left[Q_{\text {total }}\right]$ as,

$$
\mathbb{E}[Q]=\frac{\mathbb{E}\left[Q_{\text {total }}\right]}{\mathbb{E}[V]},
$$

as $\mathbb{E}[V]$ equals the mean number of vehicles in the system.

### 5.3.4.3 Packet loss

There are two types of packet loss. Foremost, packets may not be able to enter the buffer if the vehicle's buffer is full upon arrival. In addition, packets are lost if
the vehicle leaves the transmission range of the RSU, all packets at the RSU being discarded if this occurs.

There are Poisson arrivals with rate $\lambda$ for all vehicles in range with non-full buffers. Hence the total effectively accepted packet arrival rate equals,

$$
\gamma_{\text {acc,total }}=\lambda \sum_{\mathbf{x} \in \mathcal{K}} \sum_{n=0}^{C-1} \sum_{m=1}^{M} \pi(\mathbf{x}) x_{(n, m)}
$$

as in state $\mathbf{x}$, there are $\sum_{n=0}^{C-1} \sum_{m=1}^{M} x_{(n, m)}$ vehicles that have non-full buffers. The effective arrival rate at an vehicles, then relates to the total effective arrival rate as,

$$
\gamma_{\mathrm{acc}}=\frac{\gamma_{\mathrm{acc}, \text { total }}}{\mathbb{E}[V]}
$$

Moreover, as packets are either accepted or rejected upon arrival, the rejection rate $\gamma_{\text {rej }}$ equals,

$$
\gamma_{\mathrm{rej}}=\lambda-\gamma_{\mathrm{acc}}
$$

We now focus on the number of packets that are actually transmitted. Let $\gamma_{\text {tr,total }}$ denote the average departure rate from all vehicles in the system. For each system state, summing the departure rates of all vehicles, we find,

$$
\gamma_{\mathrm{tr}, \text { total }}=\sum_{\mathbf{x} \in \mathcal{K}} \sum_{n=1}^{C} \sum_{m=1}^{M} \pi(\mathbf{x}) x_{(n, m)} \mu_{(m, n)}(\mathbf{x})
$$

The transmission rate for a single vehicle $\gamma_{\mathrm{tr}}$, then relates to the total transmission rate as,

$$
\gamma_{\mathrm{tr}}=\frac{\gamma_{\mathrm{tr}, \text { total }}}{\mathrm{E}[V]}
$$

Finally as packets that are accepted are either transmitted or discarded, packets are discarded with rate,

$$
\gamma_{\mathrm{dis}}=\gamma_{\mathrm{acc}}-\gamma_{\mathrm{tr}}
$$

### 5.3.4.4 Packet delay

Another meaningful metric of the RSU is the mean packet delay $\mathbb{E}[D]$. The latter is directly related to the mean system content by Little's theorem,

$$
\begin{equation*}
\mathbb{E}[D]=\frac{\mathbb{E}\left[Q_{\text {total }}\right]}{\gamma_{\text {acc,total }}}, \tag{5.8}
\end{equation*}
$$

where $\gamma_{\text {acc,total }}$ is the effective effective arrival rate at the RSU as introduced above. Note however that the mean delay also includes the "delay" of packets that are discarded when the vehicle leaves the coverage area of the RSU. Numerical results reveal that the mean delay calculations above approximate the mean delay of packets that are effectively transmitted well, provided that the arrival rate $\lambda$ is not very high.

### 5.4 Numerical results

We now illustrate our results by means of some numerical examples. We first focus on the probability that vehicles that arrive in the coverage area of the RSU can also connect. In view of the results of Section 5.3.4.1 the connection probability can be investigated by studying the reduced Markov chain that only tracks the position of the connected vehicles.

We assume that the coverage area of the RSU is 2 km and is divided into 3 regions, the first and last region being twice as long as the middle region. That is, the first and last region are 800 m , while the middle region is 400 m . The zones are referred to as entry, production and exit zone, respectively. In Figure 5.5 the mean number of connected vehicles $\mathbb{E}[V]$ and the mean blocking probability $P_{b}$ are depicted versus the arrival rate of new vehicles $\alpha$. Different values for the vehicle speed $\nu$ are assumed as depicted. As expected, both the mean number of connected vehicles as well as the blocking probability increases for increasing arrival rates $\alpha$. In contrast, higher vehicle speeds lead to a reduction of the vehicles (reduction of the mean and the blocking probability) as the vehicles only remain for a shorter period in the coverage area of the RSU.


Figure 5.5: Moments of the number of vehicles, and the vehicle blocking probability in a system with maximum number of customers $K=10$, number of ranges $M=3$ and average traffic velocities $30,60,90$ and $120 \mathrm{~km} / \mathrm{h}$.

We now fix the speed to $90 \mathrm{~km} \mathrm{~h}^{-1}$ and retain the assumptions on the coverage area. Notice that at $90 \mathrm{~km} \mathrm{~h}^{-1}$ we have $1 / \beta_{1}=1 / \beta_{3}=32 \mathrm{~s}$ and $1 / \beta_{2}=16 \mathrm{~s}$. We consider various performance measures related to packet delivery, and therefore introduce the following assumptions on packet arrivals and packet transmissions. The vehicle arrival rate is fixed to $\alpha=0.1 \mathrm{~s}^{-1}$, meaning that there is a new connection every 10 s on average. This value corresponds to a range of a "reasonable free-flow" scenario, representing traffic flow of medium intensity (see Figures 3 and 4, and Table II of [9]). This choice is motivated by the main applicability of the RSU: RSU's in remote locations in conditions with insufficient urban infrastructure and cellular connectivity where the typical road traffic load is

| Number of zones | $M=3$ |
| :---: | :---: |
| Expected time spend <br> in the entry and exit zones | $1 / \beta_{1}=1 / \beta_{3}=32 \mathrm{~s}$ |
| Expected time spend <br> in the production zone | $1 / \beta_{2}=16 \mathrm{~s}$ |
| Transmission rate in <br> in 1st, 2d and 3d zones | $r_{t r}=[1,10,1] \mathrm{Mbit} / \mathrm{s}$ |
| Car interarrival time | $1 / \alpha=10 \mathrm{~s}$ |
| Max customers in service | $K=10$ |
| Buffer size | $C=3$ packets |
| Mean packet size | $\theta=125 \mathrm{kB}$ |

Table 5.1: Characteristics of the RSU and the traffic flow
low. Again following [9], we assume that the throughput in the production zone is $r_{t r}(2)=10 \mathrm{Mbps}$ while it is only $r_{t r}(1)=r_{t r}(3)=1 \mathrm{Mbps}$ in both entry and exit zones. Finally, we assume that at most $K=10$ vehicles can simultaneously connect to the RSU, and that the RSU stores at most $C=3$ packets for every vehicle, the mean packet size being $\theta=125 \mathrm{kB}$. For convenience, the parameters of the RSU and the traffic flow characteristics are summarised in Table 5.1

We assume that the bandwidth in each zone is shared by the vehicles within that zone that have packets waiting at the RSU. Recalling that there are $x_{(n, m)}$ packets in zone $m$ with $n$ packets, we can express the transmission rate for each vehicle as follows,

$$
\mu_{(n, m)}(\mathbf{x})=\frac{r_{t r}(m) / \theta}{\sum_{\ell=1}^{C} x_{(\ell, m)}}
$$

the total available departure rate in zone $m$ being $r_{t r}(m) / \theta$.
With the parameter values introduced above, Figures 5.6, 5.7, 5.8, and 5.9 depict various performance metrics versus the data arrival rate $\lambda$. Namely, Figures 5.6 and 5.7 depict the mean queue content and the mean packets delay, while Figures 5.8 and 5.8 show discarding and rejection rates correspondingly. In all plots, we express $\lambda$ in terms of the amount of data, rather than in terms of the number of packets. This rate is obtained by multiplying the rate in terms of the number of packets with the mean packet size. Recall that a packet is rejected if there is no room to store the packet upon arrival, while an accepted packet is discarded if it is not transmitted while the vehicle is in the coverage area of the RSU. To establish the accuracy of the perturbation approach, we depict the approximations around the points $\lambda_{0}=0.1 \mathrm{Mbit} / \mathrm{s}, \lambda_{1}=1 \mathrm{Mbit} / \mathrm{s}$ and $\lambda_{2}=5 \mathrm{Mbit} / \mathrm{s}$, corresponding to situations with low, medium and high load. In addition, we consider different orders $N$ of the perturbation on separate plots and compare with simulation results to verify the accuracy of the approximations.


Figure 5.6: Mean queue content versus the traffic load of the RSU serving at most $K=10$ vehicles in $M=3$ zones with queue capacity $C=3$ packets per vehicle.


Figure 5.7: Mean delay time versus the traffic load of the RSU serving at most $K=10$ vehicles in $M=3$ zones with queue capacity $C=3$ packets per vehicle.


Figure 5.8: Rate of discarded packets versus traffic load for the RSU serving at most $K=10$ vehicles in $M=3$ zones, with queue capacity $C=3$ packets per vehicle.


Figure 5.9: Rate of rejected packets versus traffic load for the RSU serving at most $K=10$ vehicles in $M=3$ zones, with queue capacity $C=3$ packets per vehicle.

For increasing $\lambda$, we observe an increase of the mean queue content and the packet rejection and discarding rates. If there are more arrivals, more arrivals cannot be accommodated in the buffer and are dropped. In addition, there are more packets in the buffer when the vehicle leaves the coverage area which explains the increase of the discarding rate. The mean queue content converges to the queue capacity $C$ for $\lambda \rightarrow \infty$, not entirely unexpected. The discarding rate also converges to a fixed value for $\lambda \rightarrow \infty$. This value can be calculated by noting that for $\lambda \rightarrow \infty, C$ packets are dropped when the vehicle leaves. As each vehicle remains for 80 s , the discarding rate converges to $C /(80 \mathrm{~s})=0.0375 \mathrm{~s}^{-1}$. Finally, the mean delay first increases and then again decreases again for increasing $\lambda$, the maximum delay being found for $\lambda \approx 1 \mathrm{Mbit} / \mathrm{s}$. This can be explained by the observation that when $\lambda$ initially increases, there is more buffering in all zones, leading to an increase of the delay. When $\lambda$ further increases, the delays in the entry and exit zones do not further increase as the queue is mostly full already. However, when $\lambda$ increases, more packets are transmitted in the production zone, each such transmission being 10 times as fast as transmissions in the entry and exit zones, which explains the decrease of the mean packet delay.

We now discuss the accuracy of the approximations. Comparing the approximations with the simulation results reveals that the approximations are only accurate in a region around the point where the perturbation is taken, a better approximation being obtained if one increases the order $N$ of the perturbation. While taking more terms in the approximation initially improves the region where the approximation is accurate, a further increase not extend the region. This is not unexpected as the region of convergence of the Taylor series expansions is limited by the complex poles of the stationary probabilities (as a function of the parameter). While a single perturbation does not allow for accurately assessing the performance in the complete depicted range of the $\lambda$-values, the plots clearly show that a few perturbations with overlapping convergence regions are sufficient to get accurate estimates for all $\lambda$ up to $10 \mathrm{Mbit} / \mathrm{s}$.

Finally, recall that the delay calculations also included the time discarded packets spend in the queue. To investigate the difference between the mean delay of transmitted packets $\mathrm{E}[\hat{D}]$ and the mean delay of all packets $\mathbb{E}[D]$ as calculated before, Figure 5.10 depicts both. It is observed that the results are equal for small $\lambda$, while the difference between both is limited for larger $\lambda$.

### 5.5 Conclusions

We considered a Markovian model for assessing the performance of a downlink drive-thru scenario. Unlike existing performance evaluation techniques that mainly rely on simulation results, we tackled the problem by numerical solution methods. The network model was represented by a multidimensional Markovian


Figure 5.10: Average delay time of all packets accepted for transmission $\mathbb{E}[D]$ and average delay successfully transmitted packets $\mathbb{E}[\hat{D}]$
process. Any simplifying assumptions (e.g. dividing the coverage in zones of equal transmission rate, modelling arrival process of vehicles as a Poisson process) are in line with recent analytical studies of C-ITS and drive-thru in particular. In contrast to other models, ours accounts for natural system limitations as limited buffer capacity and a limited number of customers in service. A numerical evaluation technique based on Taylor series expansion was proposed to approximate the stationary distribution of the model at hand, which combines solution speed and accuracy. Simulation results showed that the obtained solution exhibits accurate approximation for a wide range of traffic loads.

## References

[1] E. Uhlemann. Introducing connected vehicles [connected vehicles]. IEEE Vehicular Technology Magazine, 10(1), 23-31, 2015.
[2] IEEE Std. 802.11-2012, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. IEEE Std. 802.11-2012, IEEE, 2012.
[3] - IEEE Guide for Wireless Access in Vehicular Environments (WAVE) - Architecture. IEEE Std. 1609.0-2013, IEEE, 2013.
[4] Intelligent Transport Systems (ITS); Communications Architecture. ETSI EN 302665 V1.1.1, 2010.
[5] K. Sjoberg, P. Andres, T. Buburuzan,A. Brakemeier. Cooperative Intelligent Transport Systems in Europe: Current Deployment Status and Outlook. IEEE Vehicular Technology Magazine, 12(2):89-97, 2017.
[6] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, Y. Zhou. Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions. IEEE Communications Surveys \& Tutorials 17(4):2377-2396, 2015.
[7] M. Shafi, A. Molisch, P. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, G. Wunder. 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment and Practice. IEEE Journal on Selected Areas in Communications, 35(6):1201-1221, 2017.
[8] ETSI TS 102638 V1.1.1 (2009-06) Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Definitions.
[9] H. Zhou, B. Liu, F. Hou, T.H. Luan, N. Zhang,L. Gui, Q. Yu, X.S. Shen. Spatial coordinated medium sharing: optimal access control management in drive-thru internet. IEEE Transactions on Intelligent Transportation Systems, 16(5):2673-2686, 2015.
[10] M. Xing, J. He, L. Cai. Maximum-utility scheduling for multimedia transmission in drive-thru Internet. IEEE Transactions on Vehicular Technology, 65(4):2649-2658, 2016.
[11] R.F. Atallah, M.J. Khabbaz, C. M. Assi. Modeling and Performance Analysis of Medium Access Control Schemes for Drive-Thru Internet Access Provisioning Systems. IEEE Transactions on Intelligent Transportation Systems 16(6):3238-3248, 2015.
[12] C. Campolo, A. Molinaro,A. Vinel. To switch or not to switch: Service discovery and provisioning in multiradio V2R communications. In Proceedings of the 8th International Workshop on Resilient Networks Design and Modeling, IEEE, pages 281-287, 2016.
[13] C. Campolo, A. Molinaro,A. Vinel, N. Lyamin, M. Jonsson. Service discovery and access in vehicle-to-roadside multi-channel VANETs. In Proceedings of International Conference on Communication Workshop, IEEE, pages 2477-2482, 2015.
[14] E. Altman, K.E. Avrachenkov, R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. Advances in Applied Probability 36(3):839-853, 2004.
[15] J.B. Lasserre. A formula for singular perturbations of Markov chains. Journal of Applied Probability 31(3):829-833, 1994.
[16] K.E. Avrachenkov, J.A. Filar, P.G. Howlett. Analytic perturbation theory and its applications. SIAM, 2013.
[17] W.B. van den Hout. The power-series algorithm: a numerical approach to Markov processes. PhD Thesis. Tilburg University, 1996.
[18] G. Koole. On the power series algorithm. CWI, 1994.
[19] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In Performance Evaluation of Computer and Communication Systems, pages 53-80, 1993.
[20] J.P.C. Blanc, R.D. van der Mei. Optimization of polling systems with Bernoulli schedules. Performance Evaluation 22(2):139-158, 1995.
[21] B. Błaszczyszyn, T. Rolski, V. Schmidt. Advances in Queueing: Theory, Methods and Open Problems, chapter Light-traffic approximations in queues and related stochastic models. CRC Press, Boca Raton, Florida, 1995.
[22] I. Kovalenko. Rare events in queueing theory. A survey. Queueing systems 16(1):1-49, 1994.
[23] E. Evdokimova, S. Wittevrongel and D. Fiems. A Taylor Series Approach for Service-Coupled Queueing Systems with Intermediate Load. Mathematical Problems in Engineering, Article ID 3298605, 2017.
[24] E. De Cuypere, K. De Turck, D. Fiems. A Maclaurin-series expansion approach to multiple paired queues. Operations Research Letters, 42(3):203207, 2014.
[25] E. Evdokimova, K. De Turck, S. Wittevrongel and D. Fiems. Efficient Performance Evaluation of Wireless Networks with Varying Channel Conditions. In Proceedings of the 22nd International Conference on Analytical and Stochastic Modelling Techniques and Applications, pages 59-72, Springer LNCS vol. 9081, 2015.
[26] A.T. Benjamin, J.J. Quinn. Proofs that really count: The art of combinatorial proof. Mathematical Association of America, Washington, DC, 2003. ISBN 0-88385-333-7.
[27] D.M. Young, Jr. Iterative Solution of Large Linear Systems. Academic Press, 1971.

## 6

## Conclusions

This dissertation addressed multidimensional queueing models, a particular type of Markovian queueing systems joining several queues in a single model. Multidimensionality of the queueing models rises from the applications where arrivals are stored in several buffers with interdependent queueing dynamics. As queues have impact on each other, they cannot any longer be considered as independent stochastic processes and must be studied within a joint model. When analysing such systems the typical problem is associated with state space size being an exponential function of the number of queues. It leads to the so-called state space explosion problem, meaning rapid growth of the state space to the size where standard solutions are no longer applicable. For such systems this dissertation proposes effective performance evaluation methods relying on numerical techniques. In this chapter we summarise the scope of results and observations collected in this dissertation and further discuss the possible future research work.

### 6.1 Overview of the main results

- We considered various multidimensional Markovian queueing models in applications to the following fields: industrial assembly systems and multi-user downlink communication within a wireless network.
- We successfully applied Taylor and Maclaurin series expansion techniques to approximate the invariant distribution of multidimensional Markovian queueing systems. Maclaurin series approximation covers the analysis of
the light system load and overload approximation, while Taylor series serves to analyse the system behaviour under medium traffic. Terms of the expansion were calculated successively. All numerical results were depicted in terms of various performance measures and confirmed by simulations.
- Under extreme loads, both light and overload, in the framework of regular perturbation the transition rate matrix can be decomposed such that the preserved terms possess a triangular structure. This helps in solving the systems of linear equations for the terms of the Maclaurin expansion at linear complexity, by calculating elements of the terms in lexicographical or reverse lexicographical order for overload and light load respectively.
- Exploration of structural properties of the multidimensional queueing models showed that the main structural property that can be practically used to provide the performance analysis is sparsity. Transition rate matrices at hand were shown to be sparse under given model assumptions. Iterative methods, in particular Jacobi, GaussSeidel and Successive Over-Relaxation, are employed to evaluate the terms of the power series.


### 6.2 Future work

In this dissertation, applications in assembly systems and wireless communications motivated the study of multidimensional queueing models. The developed queueing models and numerical techniques, however, exhibit the flexibility to model and analyse various systems under a wide range of model assumptions. For example, while the introduced models mainly focus on single-server scenarios, the methods can easily be extended to multiserver queueing systems. Thus, the models for single access point transmission in Chapters 4 and 5 can be extended to the case of a mobile communication network served by several base stations.

The numerical analysis in this dissertation based on Maclaurin series approximates the stationary distribution of the model at hand in the region of some parameter of the system. Typically this parameter describes the system load. Results in Chapters 2 and 4 confirm that the terms of the Maclaurin series can be calculated at linear computational complexity $O(N)$ for state space size $N$. The numerical analysis yields a correct approximation. The region of accuracy is, however, limited to light traffic and overload. In order to extend the solution to intermediate loads, we proposed a Taylor series expansion in the region of intermediate load. In Chapters 3 and 5 we demonstrate fair accuracy of the approach. However, the complexity of the calculation procedure considerably exceeds the complexity for the Maclaurin series terms. Therefore, possible future work might aim to improve the procedure for calculating the Taylor series terms. In order to further improve the analysis of multidimensional queueing systems, an interesting future research
problem is to develop an approach to approximate the steady-state distribution in the entire parameter region at reasonable computational cost.

