Analyse van wachtlijnmodellen met groepsbediening

Analysis of Queueing Models with Batch Service

Dieter Claeys

Promotoren: prof. dr. ir. H. Bruneel, prof. dr. ir. J. Walraevens
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 2011 - 2012

UNIVERSITEIT
GENT

ISBN 978-90-8578-462-3
NUR 919, 992
Wettelijk depot: D/2011/10.500/66

Dankwoord

Hierbij wens ik mijn dank te betuigen aan iedereen die rechtstreeks of onrechtstreeks bijgedragen heeft bij het verwezenlijken van mijn doctoraat. In eerste instantie denk ik daarbij aan mijn promotor Herwig Bruneel, die mij de kans gegeven heeft om aan dit doctoraat te werken. Ook wil ik mijn dank uitdrukken voor de steun en het begrip die ik al die tijd gekregen heb, zowel in goede als in mindere momenten.

Daarnaast heb ik ook veel te danken aan mijn vertrouwde raadgevers Joris Walraevens, Bart Steyaert en Koenraad Laevens voor het grondig nalezen van mijn papers en inzichten te verschaffen wanneer ik vastzat met een wiskundige analyse. Ook wens ik de rest van mijn collega's te bedanken voor de aangename momenten op het werk, tijdens de middaglunch of op conferentie.

Uiteraard wens ik ook mijn dank te betuigen aan mijn vrouw Eline, mijn ouders en de rest van de familie voor hun morele steun.

Gent, november 2011
Dieter Claeys

Table of Contents

List of Figures

List of Tables

Notations and Acronyms

- A_k: number of arrivals during time slot k
- A: number of arrivals during a random slot
- $\Re_A$: radius of convergence of $A(z)$
- $\text{Arg}(z)$: principal value of the argument of z, i.e. a mapping in the interval $]-\pi,\pi]$
- β: probability that an available server initiates a new service when less customers are present than the service threshold l
- c: server capacity
- D-BMAP: discrete-batch Markovian arrival process
- $\delta\langle.\rangle$: Kronecker delta function $\delta\langle x=y\rangle=1 \iff x=y$ and $\delta\langle x=y\rangle = 0 \iff x\neq y$
- ε_i: i-th complex c-th root of 1: $\varepsilon_i \triangleq e^{\imath 2i\pi/c}$
- $\imath$: the imaginary unit
- IID: independent and identically distributed
- l: service threshold
- λ: mean number of arrivals per slot, i.e. $\lambda \triangleq \text{E}\left[A\right]$
- mod: modulo operator
- PGF: probability generating function
- ρ: load of the queueing system, defined as $\rho \triangleq \lambda \text{E}\left[T_c\right]/c$
- z_i, $0\leq i\leq c-1$: the c zeroes of $z^c - T_c(A(z))$ in the closed complex unit disk $\{z\in\mathbb{C} : |z|\leq 1\}$
- T_n: duration of the service period given that n customers are served

- $\Re_n$: radius of convergence of $T_n(A(z))$
- $\Re \triangleq \min\{\Re_n : 0 \leq n \leq c\}$
- $\Re_{T_n}$: radius of convergence of $T_n(z)$
- $\Re_T \triangleq \min\{\Re_{T_n} : 0 \leq n \leq c\}$
- $\lfloor . \rfloor$: floor function, i.e. $\lfloor x \rfloor \triangleq \max\{n \in \mathbb{Z} : n \leq x\}$
- $\lceil . \rceil$: ceil function, i.e. $\lceil x \rceil \triangleq \min\{n \in \mathbb{Z} : n \geq x\}$

Samenvatting

Dit doctoraat is het resultaat van het onderzoek dat ik gedurende vijf jaar heb verricht bij de SMACS onderzoeksgroep (Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent) en omvat de analyse van wachtlijnmodellen met groepsbediening. Een wachtlijnmodel is in feite een wiskundige abstractie van een situatie waarbij klanten aankomen en een wachtlijn vormen totdat ze bediend worden. Dergelijke fenomenen zijn alomtegenwoordig in het dagelijkse leven: mensen die wachten aan een loket in een postkantoor of bank, mensen in de wachtzaal van de dokter, vliegtuigen die wachten om te kunnen opstijgen, mensen die wachten totdat ze verbinding krijgen met iemand van het callcenter, datapakketten die opgeslagen worden in een buffer totdat het transmissiekanaal vrij is, enzovoort.

De analyse van een wachtlijnmodel vormt het onderwerp van de toegepast wiskundige discipline genaamd wachtlijntheorie en komt neer op het beantwoorden van vragen zoals: “Hoeveel klanten staan gemiddeld te wachten?”, “Hoelang moeten klanten gemiddeld wachten?”, “Is daar een grote variantie op?”, “Wat is de kans dat datapakketten verloren gaan vanwege een volle buffer?”, “Wat is de kans dat een klant extreem lang moet wachten?”, enzovoort. In de wachtlijntheorie wordt de kans dat een grootheid zoals het aantal klanten of de wachttijd zeer groot of lang is een staartprobabiliteit genoemd.

Het specifieke aan de modellen die we onderzoeken in dit doctoraat is dat het mogelijk is om klanten in groep te bedienen, met andere woorden meerdere klanten kunnen tegelijkertijd bediend worden. Een lift kan men beschouwen als een schoolvoorbeeld, aangezien meerdere mensen tegelijkertijd naar een andere verdieping gebracht kunnen worden. Ook in productie- of transportprocessen komt het dikwijls voor dat meerdere goederen samen verwerkt of getransporteerd worden. In kwaliteitscontrole bijvoorbeeld, kan classificatie van items als goed of slecht vaak efficiënter gebeuren door deze in groep in plaats van individueel te testen. Als het resultaat van een groeptest goed weergeeft, kan besloten worden dat alle items in deze groep goed zijn. In het andere geval, zijn er één of meerdere items slecht en kan men om de slechte items te localiseren de groep opsplitsen in kleinere groepen en die daarna opnieuw testen. De groeptest strategie komt vooral tot recht wanneer het percentage van slechte

items laag is.
Daarnaast worden in netwerken vaak pakketten met dezelfde bestemming en quality of service (QoS) vereisten gegroepeerd in zogenoemde bursts en worden deze bursts verstuurd over het netwerk. Dit komt de efficiëntie ten goede, omdat slechts één header per burst dient geconstrueerd te worden, in plaats van één header per individueel pakket, wat dus leidt tot een toegenomen throughput. Enkele voorbeelden van technologieën die gebruik maken van pakket aggregatie zijn Optical burst switched (OBS) netwerken en IEEE 802.11n wireless local area netwerken (WLANs).

Een inherent aspect van wachtlijnmodellen met groepsbediening is dat eenmaal de bediening gestart is, er geen nieuwe klanten meer aan de bediening kunnen toegevoegd worden, maar moeten wachten totdat de huidige bediening voorbij is, zelfs wanneer er nog plaats is in het bedieningsstation. Zo zullen mensen die toekomen wanneer de lift vertrokken is, moeten wachten totdat de lift de inzittende mensen naar hun gewenste verdieping gebracht heeft en teruggekomen is, wat mogelijk lang kan duren in hoge gebouwen. Daarom is het bij groepsbediening van belang om een doordachte beslissing te nemen wanneer de bedieningsentiteit beschikbaar is en er minder klanten aanwezig zijn dan er in principe bediend kunnen worden. Deze beslissing wordt bedieningspolitiek genoemd.
Er bestaat een uitgebreid gamma aan bedieningspolitieken. De bedieningseenheid kan bijvoorbeeld telkens wanneer deze opnieuw beschikbaar is een nieuwe bediening starten. Hoewel de reeds aanwezige klanten van deze aanpak profiteren, wordt capaciteit verspild: klanten die vlak daarna toekomen kunnen niet toegevoegd worden aan de reeds begonnen bediening. Een alternatief voor deze politiek is dat de bedieningseenheid de bediening uitstelt totdat het aantal aanwezige klanten de bedieningscapaciteit bereikt heeft, wat op zijn beurt een negatief effect heeft op de wachttijd van de reeds aanwezige klanten. Een soort van middenweg wordt geboden door de drempelgebaseerde bedieningspolitiek. Een nieuwe bediening wordt pas gestart van zodra het aantal aanwezige klanten een zekere drempelwaarde bereikt heeft. Toch is het van belang te realiseren dat zelfs met deze strategie, de wachttijden van de reeds aanwezige klanten hoog kunnen oplopen. Daarom combineren wij in deze doctoraatsthesis een drempelgebaseerde strategie met een tijdsmechanisme om te verhinderen dat klanten excessieve wachttijden ondervinden door het te lang uitstellen van een nieuwe bediening.

De bedoeling van dit doctoraat is om een uitgebreid spectrum van performantiematen te berekenen waarmee een brede waaier van situaties met groepsbediening kunnen geëvalueerd worden en waarmee men in staat is een efficiënte bedieningspolitiek te selecteren. De bestudeerde performantiematen zijn momenten, zoals de gemiddelde waarde en variantie, en staartprobabiliteiten van het aantal klanten (het aantal klanten in de wachtlijn wordt vaak bufferbezetting genoemd) en hun wachttijd.

Dit doctoraat is als volgt ingedeeld. In het eerste hoofdstuk motiveren we ons werk en introduceren we enkele cruciale begrippen zoals probabiliteitsgenererende functies (PGFs), wiens handige eigenschappen vaak aangewend worden gedurende de analyse. Daarna bestuderen we momenten en staartprobabiliteiten van de bufferbezetting in hoofdstuk 2. De resulterende formules bevatten nog onbekende probabiliteiten die numeriek berekend moeten worden. Aangezien die in sommige gevallen voor moeilijkheden kunnen zorgen, stellen we in hoofdstuk 3 benaderingen op voor de bufferbezetting. Vervolgens worden in hoofdstuk 4 de momenten en in hoofdstuk 5 de staartprobabiliteiten van de wachttijd behandeld. Om de momenten te bekomen, vatten we de wachttijd op als de som van twee niet overlappende delen, terwijl het voor de staartprobabiliteiten handiger is om de wachttijd te interpreteren als het maximum van twee tijdsperiodes. Verder vertoont het aankomstproces van klanten in de praktijk dikwijls enige vorm van afhankelijkheid, ook nog correlatie genoemd: wanneer bijvoorbeeld recent veel klanten aangekomen zijn, komen er waarschijnlijk kort daarna ook veel toe, aangezien dit kan wijzen op een piekmoment. Daarom onderzoeken we in hoofdstuk 6 de invloed van correlatie in het aankomstproces op het gedrag van groepsbedieningsfenomenen en de selectie van een efficiënte bedieningspolitiek. Tenslotte worden de belangrijkste bijdragen samengevat in het afsluitende hoofdstuk 7.

Summary

This dissertation is the result of my research work at the SMACS research group (Department of Telecommunications and Information Processing, Ghent University) and it concerns the analysis of queueing models with batch service. A queueing model basically is a mathematical abstraction of a situation where customers arrive and queue up until they receive some kind of service. These phenomena are omnipresent in real life: people waiting at a counter of a post office or bank, people in the waiting room of a doctor, airplanes waiting to take off, people waiting until they get connected with the call center, data packets which are temporarily stored into a buffer until the transmisssion channel is available, et cetera.

The analysis of queueing models constitutes the subject of the applied mathematical discipline called queueing theory and amounts to answering questions such as "How many customers are waiting on average?", "How long do customers have to wait?", "Is there a large variation on the waiting time?", "What is the probability that data packets are lost due to a full buffer?", "What is the probability that a customer suffers a lengthy delay?", et cetera. In queueing theory, the number of customers and their waiting time are often denominated by respectively buffer content and customer delay. In addition, the probability that a quantity, such as the buffer content or customer delay, is very large or lengthy, is generally called a tail probability.

The models we investigate throughout this dissertation have in common that customers can be served in batches, meaning that several customers can be served simultaneously. An elevator can be viewed as a classic example, as several people can be transported simultaneously to another floor. Also, in a variety of production or transport processes several goods can be processed together.

Furthermore, in quality control, classification of items as good or bad can often be achieved more economically by examining the items in groups rather than individually. If the result of a group test is good, all items within it can then be classified as good, whereas one or more items are bad in the opposite case, where the items can then be retested by considering smaller groups. Group testing is especially of importance when the percentage of bad items is small.

In addition, in telecommunications networks, packets with the same destination and quality of service (QoS) requirements are often aggregated into so-called bursts and these bursts are transmitted over the network. This is mainly done for efficiency reasons, since only one header per aggregated burst has to be constructed, instead of one header per single information unit, thus leading to an increased throughput. Technologies using packet aggregation include for instance Optical burst switched (OBS) networks and IEEE 802.11n wireless local area networks (WLANs).

An inherent aspect of batch service is that newly arriving customers cannot join the ongoing service, even if there is free capacity (we denominate the maximum number of customers that can be served simultaneously by server capacity). For instance, an arriving person cannot enter an elevator that has just left, even if space is available. This person has to wait until the elevator has transported its occupants to their requested floors and has returned, which might take a long time in high buildings. In view of this, it is of importance to take a well-considered decision when the server becomes available and finds less customers than it can serve in theory. This decision is called the service policy.

A whole spectrum of service policies exist. The server could, for instance, start serving the already present customers immediately. Although the present customers benefit from this approach, capacity is wasted: customers that arrive later cannot join the ongoing service. An alternative for this so-called immediate-batch service policy is the full-batch service policy. In this case, the available server postpones service until the number of present customers reaches or exceeds the server capacity, which, in turn, has a negative effect on the delay of the customers waiting to form a full batch (postponing delay). The threshold-based policy is a kind of compromise between immediate-batch service policy and full-batch service policy. When the number of present customers is below some service threshold, service is postponed, whereas service is initiated when the number of present customers reaches or exceeds this threshold. It is important to realize that even with this compromise, long postponing delays are possible. Therefore, in this dissertation, we combine a threshold-based policy with a timer mechanism that avoids excessive postponing delays.

The purpose of this dissertation is to calculate a large spectrum of performance measures, which enable to evaluate a broad set of situations with batch service and aid in selecting an efficient service policy. The studied performance measures are moments, such as the mean value and variance, and tail probabilities of the buffer content and the customer delay.

This dissertation is structured as follows. In chapter 1, we motivate our work and we introduce crucial concepts such as probability generating functions (PGFs), whose useful properties are frequently relied upon throughout the analysis. Then we deduce moments and tail probabilities of the buffer content

in chapter 2. The resulting formulas still contain unknown probabilities that have to be calculated numerically. As this might become unfeasible in some cases, we compute in chapter 3 approximations for the buffer content. Next, moments and tail probabilities of the customer delay are covered in respectively chapters 4 and 5. In order to analyze the moments, we conceive the customer delay as the sum of two non-overlapping parts, whereas for the tail probabilities, it turns out to be more convenient to interpret the delay as the maximum of two time periods. Further, in real life the customer arrival process often exhibits some kind of dependency. For instance, if a large amount of customers have recently arrived, it is likely that many customers arrive in the near future, as it might be an indication of a peak moment. Therefore, we investigate in chapter 6 the influence of dependency in the arrival process on the behaviour of batch-service phenomena and on the selection of an efficient service policy. Finally, the main contributions are summarized in chapter 7.

Chapter 1

Introduction

1.1 Queueing theory

Queueing systems are omnipresent in daily life: people waiting at a counter of a post office or bank, people in the waiting room of the doctor, airplanes waiting to take off, people waiting until they get connected to the call center, traffic jams, et cetera, are all situations whereby customers (people, airplanes, cars, et cetera) queue up until they receive some kind of service (get money, being examined by the doctor, et cetera). Also, in telecommunications, there are numerous occasions where packets (customers) are stored in a queue until the transmission channel (the server) is available. In practice, it is often of importance to study the behaviour of queueing systems. For instance, the manager of a call centre has to assess the efficiency of the centre in order to decide whether or not to recruit additional personnel. The manager will make a decision based on the answers on several questions, among which "What is the mean number of customers waiting to get connected?", "What is the mean time that customers have to wait until they get connected?", "What is the probability that a customer hangs up because he or she has to wait too long?", et cetera.

Providing answers to such questions constitutes the subject of **queueing theory**. Basically, a **queueing model** is developed for the system under consideration, whereupon **performance measures** are deduced by which the system can be evaluated. A queueing model consists, broadly speaking, of two parts, the queue and one or more servers. Customers arrive and are stored in the queue until they are processed by a server (Fig. 1.1). After the model has been developed, several performance measures are computed. In general, *moments* (such as the mean value and variance) and *tail probabilities* (a tail probability of a random variable X is $\Pr[X = n]$ or $\Pr[X > n]$ for n a large number) of the *buffer content* (the number of customers in the queue and/or in service) and *customer delay* (time that customers have to wait) are the preeminent performance measures.

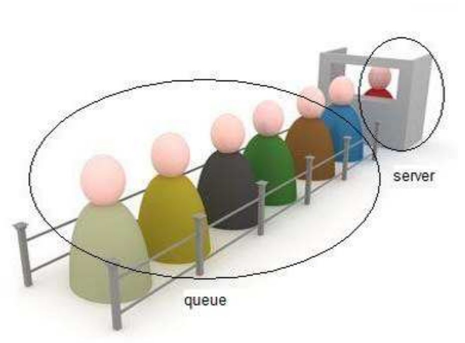


Figure 1.1: Example of a queueing system: people (customers) queue up until they are assisted by the employee at the counter (the server)

The Danish mathematician A.K. Erlang (1878-1929) is considered to be the founder of queueing theory. In 1917, he studied the holding times of conversations in telephone exchanges (his paper was translated to French in 1925 to acclaim worldwide recognition [54]). Since then, queueing theory has been applied in many disciplines, such as health care and emergency planning ([23]; [95]), transportation (car, train and air traffic congestion control, [61]; [84]), stock management and production process planning ([27]; [102]), machine breakdowns and repairs ([55]; [110]), database management and computer networks ([73]), and many others.

When studying a queueing system, it is of the utmost importance to develop an appropriate model. Several aspects have to be specified such as the frequency at which customers arrive, the number of present servers, the speed of the servers, the number of places in the waiting room, et cetera. In order to represent characteristics of queueing models in a concise manner, Kendall's shorthand notation ([72]) in the form $A/B/C/D/E$ has widely been used. Here, A and B denote interarrival- and service-time distributions (an interarrival time is the time between two consecutive instants on which one or more customers arrive, and the service time of a customer is the time required to serve that customer), C specifies the number of servers, D represents the queue size (i.e. the maximum number of customers that can be stored in the queue) and E characterises the service discipline (i.e. the order in which customers are served)[1]. Often, D and E are omitted from the notation. When D is not mentioned, the queue capacity is assumed to be infinite. If E is not specified, the service discipline is irrelevant, or it is the usual first-come-first-served (FCFS) discipline, whereby

[1] Classically, the fifth component represents the size of the population. The service discipline is then characterised by the sixth component. We have chosen not to mention the population size in the text, as it is always assumed to be infinite throughout this dissertation and in the cited papers. Moreover, we intend not to add extra complexity with definitions that do not contribute to the essence of this dissertation.

customers are served in the order they arrived in.
Some common notations for A are:

- M: memoryless, which means that the interarrival times are exponentially distributed. In addition, one customer arrives at each arrival instant, which we denominate by **single arrivals**.

- M^X: extension of M, whereby several customers (instead of only one customer) can arrive at an arrival instant (**batch arrivals** instead of single arrivals).

- Geo: the interarrival times are geometrically distributed and it concerns single arrivals.

- Geo^X: generalisation of Geo whereby more than one customer can arrive at an arrival instant. Hence, the superscript X indicates that customers can arrive in batches[2].

As the exponential distribution is continuous and the geometric distribution discrete, models with codes M and M^X for A are classified as **continuous-time** queueing models, whereas Geo and Geo^X are categorized as **discrete-time** queueing models. In case of a discrete-time model, the time-axis is divided into fixed-length contiguous time periods, called slots, and the interarrival times are expressed as an (integral) number of slots. Also, services can only be initiated and terminated at slot boundaries, whereas in the continuous-time counterpart, a new customer is served immediately when the server is available. Whenever the queueing system under investigation has a slotted nature, it is appropriate to adopt a discrete-time queueing model. For instance, in telecommunications, operations are synchronized to the system clock, so that a slot corresponds to the clock interval.
The above mentioned interarrival-time distributions have in common that the number of arrivals in some time interval is independent of the amount of arrivals in another, non-overlapping time interval, which is called **independent arrivals**. Codes such as MAP, $BMAP$ and their discrete-time counterparts $D-MAP$ and $D-BMAP$ correspond to **dependent (or correlated) arrivals**. We discuss $D-BMAP$ in chapter 6.
With respect to the distribution of the service times, we mention the following codes for B:

- M: exponential distribution.

- E_k: Erlang distribution with shape parameter k, i.e. a convolution of k exponential distributions.

- D: deterministic distribution (constant).

- G: general distribution, i.e. an unspecified distribution.

[2] In some papers, X represents the exact distribution of the number of customers in the arriving batches.

1.2 Batch service

Beside service times and number of servers, another distinction can be made in the service process: traditional versus **batch service**. Whereas a traditional server can serve one customer at a time, a batch server can process several customers simultaneously. In technical terms, a batch server processes batches which can contain up to c customers (c is called the **server capacity**).
An elevator can be viewed as a clear example of batch service, since elevators can bring several people simultaneously to another floor. Other examples include transport vehicles, busses, ship locks, ovens in production processes, attractions in amusement parks, et cetera. Furthermore, in telecommunications, it is often the case that information packets are grouped in larger entities (batches) and these batches are transmitted instead of each packet individually. This is mainly done for efficiency reasons, since only one header per aggregated batch has to be constructed, instead of one header per single information unit, thus leading to an increased throughput. Technologies using packet aggregation include Optical burst switched (OBS) networks [36], [96] and IEEE 802.11n wireless local area networks (WLANs) [80]. More applications can, for instance, be found in [18].
An inherent aspect of batch-service systems is that newly arriving customers cannot join the ongoing service, even if there is free capacity. For instance, an arriving person cannot enter an elevator that has just left, even if space is available. In view of this, a decision has to be taken, called the **service policy**, when the server becomes available and finds less than c customers in the queue. The server could, for instance, start serving the already present customers immediately ([41]; [90]; [117]). Although the present customers benefit from this approach, capacity is wasted: customers that arrive later cannot join the ongoing service. An alternative for this so-called immediate-batch service policy is the full-batch service policy ([30]; [32]). In this case, the available server postpones service until the system contains at least as many customers as its capacity, which, in turn, has a negative effect on the delay of the customers waiting to form a full batch (*postponing delay*). The threshold-based policy [65] (also called minimum batch size service policy [37], or general batch-service policy [94]) is a kind of compromise between immediate-batch service policy and full-batch service policy. When the number of present customers is below some **service threshold** l, service is postponed, whereas service is initiated when the number of customers reaches or exceeds l. In fact, immediate-batch service policy and full-batch service policy are special cases. Indeed, the first corresponds to $l = 1$, whereas $l = c$ represents the latter. It is important to realize that even with this compromise, long postponing delays are possible when $l > 1$. Therefore, in this dissertation, we combine a threshold-based policy with a timer mechanism that avoids excessive postponing delays: when l ($0 \leq l \leq c$) or more customers are present, a new service is initiated, whereas otherwise a new service is started only with probability β. Note that the threshold-based policy is a special case of this policy ($\beta = 0$). To finalize this section, we

mention that batch service is indicated in the Kendall notation by adding the superscript (l, c) to the code for B (or (l, c, β) if we include our mechanism that avoids excessive postponing delays).

1.3 Motivation

Bailey [9] was presumably the first to investigate a batch-service queueing system. He obtained the distribution of the buffer content at random time epochs in an $M/G^{(0,c)}/1$ queueing system. Since then, many papers, as well in continuous as in discrete time, have been published about batch-service. Downton [51] examined the customer delay for the same queueing model as in [9]. Neuts [91] studied the buffer content at random time epochs and at service completion times in an $M/G^{(l,c)}/1$ system. Further, the customer delay in an $M/M^{(l,c)}/1$ system was calculated by Medhi [88]. Chaudhry and Templeton [32] deduced the distributions of the buffer content at random epochs and the customer delay in the systems $M/G^{(0,c)}/1/K$, $M/G^{(0,c)}/1$, $M/G^{(c,c)}/1$ and the discrete $Geo/G^{(0,c)}/1$. Powell and Humblet [94] calculated the buffer content at service completion times in the discrete $Geo^X/G^Y/1$ system for several possible service policies Y (these are combinations of thresholds and server vacations - i.e. the server being unavailable to serve for some time [50]; [58]; [108]), and Zhao and Campbell [117] studied the buffer content at random slot boundaries in the discrete $Geo^X/D^{(1,c)}/1$ system. Next, Chang and Choi [29] investigated the buffer content at random, service termination and arrival epochs in the $Geo^X/G^{(1,Y)}/1/K$ queue with varying server capacity Y and vacations, Yi et al. [115] evaluated the buffer content at the same epochs as [29] in the $Geo^X/G^{(a,Y)}/1/K$ queue and Arumuganathan and Jeyakumar [7] examined the buffer content at various time epochs in an $M^X/G^{(l,c)}/1$ model. The buffer content at several time epochs in an $M/G^{(l,c)}/1$ system with vacations was computed by Sikdar and Gupta [103]. Kim and Chaudhry [75] studied equivalences between batch-service and multi-service (i.e. c servers of capacity one) systems and Samanta et al. [98] derived the buffer content at various time instants in a discrete $Geo^X/G^{(l,c)}/1/K$ system with vacations. Other papers concerning batch service include [3], [28], [33], [34], [63], [64], [66], [70], [104].

From the above literature overview, it follows that most research concerning batch-service queueing models has focused on the buffer content, that the customer delay has only been studied in the case of single arrivals and that nearly always independent arrivals are considered. In this dissertation, we study several novel aspects of a versatile discrete-time batch-service queueing model (this model is described in detail in section 1.5). In chapter 2, we compute a fundamental formula from which a large spectrum of known as well as new results regarding the buffer content of such batch-service queues are extracted. As in all batch-service models, the resulting formulas contain unknown probabilities that have to be calculated numerically. This can become an unfeasible assignment especially for large c. Therefore, we deduce in chapter 3 light- and heavy-traffic approximations of the buffer content that require only a few (and sometimes no) numerical calculations. To the best of our knowledge, light-and heavy-traffic approximations have not been studied before for batch-service queueing models. In chapter 4, we examine moments of the customer delay, whereas in chapter 5, we deal with tail probabilities of the customer delay. As the inclusion of batch arrivals forms the novel aspect in the delay study of a batch-service system, we compare our results to those of the corresponding single-arrival systems. It will become clear that batch arrivals have an undeniable impact and therefore have to be included in the model. Since in nearly all batch-service queueing models considered in the literature, independent arrivals are considered, we include in chapter 6 correlated arrivals in our model and we evaluate the influence of it on the behaviour of the system. Finally, we conclude this dissertation by summarizing the main contributions in chapter 7.

1.4 Probability generating functions (PGFs)

Throughout this dissertation, we extensively make use of **probability generating functions (PGFs)**. Let z be a complex variable. The PGF $X(z)$ of some random variable X (X can represent, for instance, the number of customers in the queue at some random time instant) is then defined as

$$X(z) \triangleq \mathrm{E}\left[z^X\right] = \sum_{n=0}^{\infty} \Pr\left[X = n\right] z^n \ ,$$

with $\mathrm{E}\left[.\right]$ the expectation operator. The main reason to resort to PGFs stems from their useful properties, which are mentioned below.

Analyticity

The **radius of convergence** $\Re_X$ of a PGF $X(z)$ is defined so that $X(z)$ is analytic for $|z| < \Re_X$ and not analytic for $z = \Re_X$[3]. It holds that $\Re_X \geq 1$. As a result, each PGF $X(z)$ is an analytic function of z in the open complex unit disk $\{z \in \mathbb{C} : |z| < 1\}$. In particular, $X(z)$ is bounded in the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$, implying that a PGF cannot have poles in this area. This property is frequently invoked throughout this dissertation. We often obtain a PGF as a fraction that contains unknowns in the numerator. We then prove that the denominator has several zeroes in the closed complex unit disk. On account of the analytic property of PGFs, the numerator also has to vanish in these zeroes, which leads to a set of equations by which the unknowns can be determined.

[3] Vivanti's theorem was implicitly invoked in this definition for radius of convergence. Vivanti's theorem states that if $X(z)$ is a power series with real positive coefficients (which is the case when $X(z)$ is a PGF) and with radius of convergence $\Re_X$, then $X(z)$ is not analytic at $z = \Re_X$.

Normalization condition

Every PGF $X(z)$ satisfies the equation $X(1) = \sum_{n=0}^{\infty} \Pr[X = n] = 1$. Typically, one extra equation is required in the abovementioned set of equations. The normalization condition provides this equation.

Probability generating property

Since $\Pr[X = n]$ is the coefficient of z^n in the Taylor series expansion of $X(z)$ about $z = 0$, the mass function can be extracted from the PGF by computing derivatives at $z = 0$:

$$\Pr[X = n] = \frac{1}{n!} \frac{d^n X(z)}{dz^n}\bigg|_{z=0} \ .$$

This implies in particular that

$$\Pr[X = 0] = X(0) \ .$$

Hence, it is possible to deduce the mass function via these formulas. However, it is frequently required to calculate *tail probabilities* of X, i.e. $\Pr[X > n]$ for n a large number. In this case, it becomes unfeasible to apply the probability generating property as calculating n-th order derivatives becomes unfeasible when n is large. We then resort to an approximation technique, which is based on **Darboux's theorem**:

Theorem 1. *Suppose the power series $Y(z) = \sum_{n=0}^{\infty} y(n) z^n$ with positive real coefficients $y(n)$ is analytic near 0 and has only algebraic singularities α_j on its circle of convergence $|z| = \Re_Y$, in other words, in a neighbourhood of α_j we have*

$$Y(z) \sim \left(1 - \frac{z}{\alpha_j}\right)^{-\omega_j} G_j(z) \ ,$$

i.e.

$$\lim_{z \to \alpha_j} \frac{Y(z)}{G_j(z)} \left(1 - \frac{z}{\alpha_j}\right)^{\omega_j} = 1 \ ,$$

where $\omega_j \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$ and $G_j(z)$ denotes a nonzero analytic function near α_j. Let $\omega \triangleq \max_j Re(\omega_j)$ denote the maximum of the real parts of ω_j. Then we have

$$y(n) = \sum_j \frac{G_j(\alpha_j)}{\Gamma(\omega_j)} n^{\omega_j - 1} \alpha_j^{-n} + o\left(n^{\omega - 1} \Re^{-n}\right) \ ,$$

with the sum taken over all j with $Re(\omega_j) = \omega$ and $\Gamma(\omega)$ the Gamma-function of ω (with $\Gamma(n) = (n-1)!$ for n discrete).

Note on the one hand that when $X(z)$ is a PGF corresponding to a random variable X, it holds that

$$\begin{aligned}\frac{X(z)-1}{z-1} &= \frac{\sum_{n=0}^{\infty} \Pr[X=n]\, z^n - 1}{z-1} \\ &= \sum_{n=0}^{\infty} \Pr[X=n] \frac{z^n-1}{z-1} \\ &= \sum_{n=0}^{\infty} \Pr[X=n] \sum_{k=0}^{n-1} z^k \\ &= \sum_{k=0}^{\infty} z^k \sum_{n=k+1}^{\infty} \Pr[X=n] \\ &= \sum_{k=0}^{\infty} \Pr[X>k]\, z^k \ ,\end{aligned}$$

meaning that $[X(z)-1]/(z-1)$ is a power series with positive real coefficients $\Pr[X>n]$. In addition, due to the analyticity of PGFs in the open unit disk, $[X(z)-1]/(z-1)$ is analytic near $z=0$. On the other hand, when in a neighborhood of the **dominant singularities** α_j (a dominant singularity is a singularity with smallest modulus, thus on the radius of convergence of $X(z)$)

$$X(z) \sim \left(1-\frac{z}{\alpha_j}\right)^{-\omega_j} G_j(z) \ ,$$

it also holds that

$$\frac{X(z)-1}{z-1} \sim \left(1-\frac{z}{\alpha_j}\right)^{-\omega_j} \frac{G_j(z)}{z-1} \ .$$

Application of Darboux's theorem then yields

$$\Pr[X>n] \approx \sum_j \frac{G_j(\alpha_j)}{\alpha_j-1} \frac{n^{\omega_j-1}}{\Gamma(\omega_j)} \alpha_j^{-n} \ . \tag{1.1}$$

Summarized, the approach boils down to locating the dominant singularities of the PGF $X(z)$ and then relying on formula (1.1). Throughout this dissertation, the dominant singularities are always poles, so that we deal with one of the following special cases:

Corrolary 1. *When the PGF $X(z)$ has k dominant singularities α_j, all poles with multiplicity 1, formula (1.1) transforms into*

$$\Pr[X>n] \approx \sum_{j=1}^{k} \frac{G_j(\alpha_j)}{\alpha_j-1} \alpha_j^{-n} = \sum_{j=1}^{k} \frac{\alpha_j^{-(n+1)}}{1-\alpha_j} \frac{N_X(\alpha_j)}{D'_X(\alpha_j)} \ , \tag{1.2}$$

with $N_X(z)$ and $D_X(z)$ respectively the (mutually indivisible) numerator and denominator of $X(z)$.

Corrolary 2. *When the PGF $X(z)$ has one dominant singularity, being a pole α with multiplicity m, formula (1.1) transforms into*

$$\Pr[X>n] \approx \frac{G(\alpha)}{\alpha-1} \frac{n^{m-1}}{(m-1)!} \alpha^{-n} \ . \tag{1.3}$$

Corrolary 3. *When the PGF $X(z)$ has one dominant singularity, being a pole α with multiplicity 1, formula (1.1) transforms into*

$$\Pr[X > n] \approx \frac{G(\alpha)}{\alpha - 1}\alpha^{-n} = \frac{\alpha^{-(n+1)}}{1-\alpha}\frac{N_X(\alpha)}{D_X^{'}(\alpha)} \ . \tag{1.4}$$

Moment generating property

When the radius of convergence of the PGF $X(z)$ is larger than one (which we assume for every PGF from now on), all order moments of X exist and can be calculated by taking derivatives of $X(z)$ at $z = 1$. For instance, the first moment (the mean value) is equal to:

$$\mathrm{E}[X] = X^{'}(1) \ ,$$

(we use primes to indicate derivatives) and the second moment reads:

$$\mathrm{E}[X^2] = X^{''}(1) + X^{'}(1) \ .$$

As a result, the variance becomes

$$\mathrm{Var}[X] = X^{''}(1) + X^{'}(1) - X^{'}(1)^2 \ .$$

Sum of independent random variables

The PGF $Y(z)$ of a sum of independent random variables $(X_1, X_2, ..., X_n)$ equals the product of the corresponding PGFs:

$$Y(z) = X_1(z)X_2(z)...X_n(z) \ .$$

In the probability domain, this would lead to a convolution of the mass functions of the random variables X_i, which is far from straightforward to calculate, because a convolution of n mass functions leads to $n-1$ summations:

$$\Pr[Y = k] =$$
$$\sum_{i_1=0}^{k}\sum_{i_2=0}^{k-i_1}\cdots\sum_{i_{n-1}=0}^{k-i_1-\ldots-i_{n-2}} \Pr[X_1 = i_1]\Pr[X_2 = i_2]\ldots\Pr[X_n = k - i_1 - \ldots - i_{n-1}] \ .$$

Joint PGFs

Let x and z be complex variables. The joint PGF, say $\tilde{X}(z,x)$, of two random variables X and Y is then defined as

$$\tilde{X}(z,x) \triangleq \mathrm{E}\left[z^X x^Y\right] = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\Pr[X = n, Y = m]\, z^n x^m \ .$$

From this definition, it is clear that the marginal PGFs $X(z)$ and $Y(z)$ of respectively X and Y are equal to

$$X(z) = \tilde{X}(z,1) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\Pr[X = n, Y = m]\, z^n = \sum_{n=0}^{\infty}\Pr[X = n]\, z^n \ ,$$

$$Y(z) = \tilde{X}(1,z) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \Pr[X=n, Y=m]\, z^m = \sum_{m=0}^{\infty} \Pr[Y=m]\, z^m \ .$$

This technique is often referred to as summing out a random variable. In addition, the PGF of the sum of X and Y is equal to

$$\mathrm{E}\left[z^{X+Y}\right] = \tilde{X}(z,z) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \Pr[X=n, Y=m]\, z^{n+m} \ .$$

Finally, when X and Y are statistically independent, the joint PGF of X and Y equals the product of the marginal PGFs:

$$\begin{aligned}\tilde{X}(z,x) &= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \Pr[X=n, Y=m]\, z^n x^m \\ &= \sum_{n=0}^{\infty} \Pr[X=n]\, z^n \sum_{m=0}^{\infty} \Pr[Y=m]\, x^m \\ &= X(z)Y(x) \ .\end{aligned}$$

Note that the property of the sum of independent random variables is in fact a special case of this property.

To close, we would like to stress that the definition and properties for the joint PGF of three or more random variables is analogous as those mentioned for two random variables.

1.5 Model description

This section summarizes the properties of the model under investigation.

- We consider a discrete-time queueing model, i.e. the time axis is divided into fixed-length contiguous periods, referred to as slots.
- Several customers can arrive during each slot (batch arrivals). The number of customers that arrive during slot k is denoted by A_k. With the exception of chapter 6, we consider independent arrivals[4]. In other words, we assume in chapters 2-5 that the sequence $\{A_k\}_{k\geq 1}$ consists of independent and identically distributed (IID) random variables, with common PGF $A(z)$. The number of customer arrivals during an arbitrary slot is denoted by A and has, on account of the IID character of the arrivals, PGF $A(z)$. The mean number of customer arrivals during a random slot, $\mathrm{E}[A]$, is denoted by λ and is, owing to the moment generating property of PGFs, equal to $A^{'}(1)$.
- For mathematical convenience, we assume that the queue is infinitely large. Therefore, all arriving customers can enter the queue and will

[4] The combination discrete time, batch arrivals and independent arrivals is equivalent with geometrically distributed interarrival times whereby a slot corresponds to unity length of time and whereby batches consisting of one or more customers arrive at an arrival instant.

eventually be served if the system is stable (which we assume to hold - see assumption 1 on page 13). This assumption is not as stringent as it seems, as in practice queues are very large in order to avoid that customers are lost due to a full queue.

- There is one batch server of capacity c (c fixed), which means that the server can process up to c customers simultaneously. When the server becomes available and finds at least as many customers as the service threshold l ($0 \leq l \leq c$), it initiates a new service, whereas when the amount of available customers is smaller than l, the server initiates a service with probability β and with probability $1 - \beta$ it postpones its service. This feature avoids that customers suffer excessive delays because the server waits to initiate service until enough customers have arrived. We assume that the already present customers remain in the queue when the server postpones service. Hence, during each slot, the **system content** consists of the customers being served (the **server content**) and the customers waiting in the queue (the **queue content**). Throughout this dissertation we also frequently mention the term **buffer content**, which we adopt as a generic expression for the system content, the queue content and the server content.

- A service period is the period between the start and end of the service of one batch of customers. The service periods are synchronized to slot marks, in the sense that the server always starts and ends processing at slot boundaries. As a result and because customers arrive during slots (thus not on slot marks), an arriving customer has to wait for service at least until the beginning of the next slot. This kind of synchronization is also known as LAS-DA (late arrival system with delayed access; see e.g. [98]). The remaining part of the slot wherein a customer arrives is not included in what we denominate the customer delay, since we express the customer delay as an integral number of slots.

- A service time is the length of a service period, in terms of number of slots. The service time of a batch is dependent on the number of customers within it. Given this number, the service time is independent of all previous service times. We denote the service time of a batch containing j customers by T_j and its corresponding PGF is represented by $T_j(z)$. We assume that $T_j(0) = 0$. Note that we do not demand that $T_0(z) = 1$, which means that the server might be serving a batch containing no customers. This can be conceived as a server vacation whereby the length is distributed according to $T_0(z)$. During such a server vacation, the server can do some other useful work. For instance, when an oven in a production process has no goods to process, the operator might execute some maintenance work. The maintenance has to be completed before new goods can be processed and can thus be conceived as a server vacation.

- The order at which customers are selected from the queue (the queueing discipline) is irrelevant for the buffer content (chapters 2, 3 and 6). However, the queueing discipline plays a role when studying the customer delay (chapters 4 and 5). We investigate the customer delay for the first-come-first-served (FCFS) discipline.

The shorthand notation for this model thus reads $Geo^X/G^{(l,c,\beta)}/1$. Our model includes some special cases:

- $\beta = 0$: the threshold-based service policy: when less than l customers are present, no service is initiated ([37]; [65]; [91]; [94]).

- $\beta = 0$ and $l = 1$: immediate-batch service policy: when at least one customer is present, a new service is initiated ([41]; [90]; [117]).

- $\beta = 0$ and $l = c$: full-batch service policy: the server only processes full batches, i.e. batches containing c customers ([30]; [32]; [41]).

- $l = 0$: the server always starts a new service when it becomes available, even when no customers are present ([9]; [29]; [32]; [51]).

Remark 1. *When studying buffer-related quantities, such as the system content, the queue content, et cetera, we will always observe these at slot marks. As service times are also synchronised at slot boundaries, we have to order these events to avoid ambiguity. We assume that the observation epochs are immediately after the potential service initiation epochs (and thus also after the service termination epochs). Since customers arrive during slots, but not on slot boundaries, the observation epochs occur thus prior to potential arrivals. In order to illustrate the order of possible events in a slot, we have illustrated them in Fig. 1.2.*

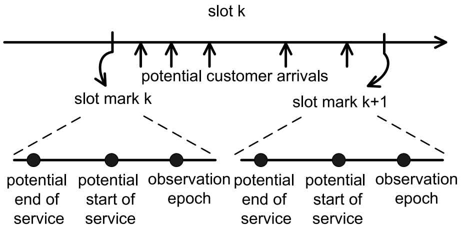


Figure 1.2: Overview of the possible events in a slot

1.6 Assumptions

The results in this dissertation are valid under the following assumptions:

Assumption 1. *The* ***load*** $\rho \triangleq \lambda \mathrm{E}\left[T_c\right]/c < 1$*. In the case that* $\rho > 1$*, more customers arrive on average than the system can process (when many customers are present, the server nearly always processes c customers, so that on average c customers leave every* $\mathrm{E}\left[T_c\right]$ *slots), which makes the system unstable. In the other case, the system is stable (see e.g. [94] for a proof), implying that after a sufficiently long time period, the system reaches a* ***steady state*** *(also called stochastic equilibrium), meaning that the distributions of all quantities such as the buffer content become independent of the slot number. In this dissertation, we focus on the steady-state behaviour of the system.*

Assumption 2. *The radius of convergence of each PGF is strictly larger than 1. This implies that all order moments are finite and can be calculated by means of the moment generating property of PGFs. As mentioned in section 1.4, we designate the radius of convergence of some random variable X by* $\Re_X$*. In addition, we define* $\Re_n$ *as the radius of convergence of* $T_n(A(z))$ *and* $\Re \triangleq \min\{\Re_n : 0 \leq n \leq c\}$ *and* $\Re_T \triangleq \min\{\Re_{T_n} : 0 \leq n \leq c\}$*.*

Assumption 3. $\Re_n \leq \Re_A$*. It is worth mentioning that we believe that this assumption is actually a fact, as we have not been able to construct one counterexample*[5]*. However, as it is tedious to prove that* $\Re_n \leq \Re_A$*, we mention it as an assumption.*

Assumption 4. $z^c - T_c(A(z))$ *is aperiodic, meaning that the highest common factor of the set of integers* $\left\{\{c\} \cup \left\{n \in \mathbb{N} : \frac{d^n}{dz^n} T_c(A(z))\Big|_{z=0} \neq 0\right\}\right\}$ *equals 1.*

Assumption 5. $\lim_{z \uparrow \Re} T_c(A(z))/z^c > 1$*. This assumption will assure that* $z^c - T_c(A(z))$ *has a zero in the interval* $]1, \Re[$ *(see e.g. [106]), which in turn entails that the tail probabilities of the studied quantities (for instance the system content at slot boundaries, the customer delay, et cetera) are dominated by this zero and not by the specific dominant singularity of* $T_c(A(z))$*. Although we thus exclude some PGFs* $T_c(A(z))$*, the commonly adopted PGFs satisfy this assumption. The main advantage is that we can present a general solution whereas otherwise an ad hoc approach would have to be adopted for each PGF* $T_c(A(z))$*.*

[5] When trying to construct a counterexample, one should verify that the constructed $A(z)$ and $T_n(z)$ are indeed PGFs, by checking the normalization condition and verifying that the coefficients in the Taylor series expansions of $A(z)$ and $T_n(z)$ about $z = 0$ are probabilities.

Chapter 2

Buffer content: exact analysis

2.1 Preface

In the introduction, we have mentioned that most research concerning batch-service queueing models has focused on some specific aspect of the buffer content. In this chapter, we compute a fundamental formula (section 2.2) - the joint PGF of the queue content, the server content and the remaining service time - from which we extract an entire gamut of known as well as new results regarding the buffer content (section 2.3). In section 2.4, we briefly mention how performance measures can be calculated from these quantities and, in section 2.5, we demonstrate that these expressions are useful tools to select a good service policy. Finally, we show how our results can be applied in the study of group-screening policies (section 2.6).

In our paper [37], we have computed the PGF of the system content in a model that is included as a special case in the model discussed throughout this dissertation: it adopts the threshold-based service policy and the service times are independent of the number of customers in the served batches (hence $\beta = 0$ and $T_n(z) = T_c(z), \ \forall n$). In [43], we have deduced the joint PGF of the queue content, the server content and the remaining service time for the same model as in [37] and we have extracted various quantities from this joint PGF. In [38], we have extended our model from [37] and [43] so that the service times become dependent on the number of served customers. In this chapter, we opt to analyze the buffer content immediately for the versatile model described in section 1.5, as the analysis runs mainly parallel as in our contributions [37], [38] and [43].

2.2 Joint PGF

In this section, we compute the steady-state joint PGF $V(z,x,y)$ of the queue content, the server content and the remaining service time of the batch in service:

$$V(z,x,y) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\right] \ ,$$

with Q_k (S_k) the queue (server) content and R_k the remaining service time at slot boundary k and z, x and y being complex variables.
We commence by writing down the system equations, which express the relation between $(Q_{k+1}, S_{k+1}, R_{k+1})$ and (Q_k, S_k, R_k):

$$(Q_{k+1},S_{k+1},R_{k+1}) = \begin{cases} (Q_k+A_k, S_k, R_k-1) & \text{if } R_k > 1 \ , \\ (0, Q_k+A_k, T_{Q_k+A_k}) & \text{if } R_k \leq 1 \text{ and } l \leq Q_k+A_k < c \ , \\ (Q_k+A_k-c, c, T_c) & \text{if } R_k \leq 1 \text{ and } Q_k+A_k \geq c \ , \\ (0, Q_k+A_k, T_{Q_k+A_k}) & \begin{array}{l}\text{if } R_k \leq 1,\ Q_k+A_k<l \text{ and service starts} \\ \text{(with probability } \beta) \ ,\end{array} \\ (Q_k+A_k, 0, 0) & \begin{array}{l}\text{if } R_k \leq 1,\ Q_k+A_k<l \text{ and service does} \\ \text{not start (with probability } 1-\beta) \ .\end{array} \end{cases}$$

Indeed, in the first case ($R_k > 1$), the ongoing service continues during slot $k+1$, so that customers that have arrived during slot k are stored in the queue (even when $0 \leq S_k < c$, because customers cannot join the ongoing service). In the other cases, the server is available at slot mark $k+1$, because $R_k \leq 1$ means that either the server was not processing during slot k ($R_k = 0$) or slot k was the last slot of a service period ($R_k = 1$). Whether a new service is initiated or not at slot $k+1$ is described by the rules mentioned in the model description (section 1.5) and is thus dependent on the number of available customers, which is equal to $Q_k + A_k$ because the customers that were in service during slot k (if any) leave the system at the end of slot k. Note that although $R_k = 0$ implies $S_k = 0$ (when there is no service, no customers are in the server) and $Q_k < l$ (when $Q_k \geq l$, a service would be initiated if the server was available), $S_k = 0$ does not necessarily imply $R_k = 0$. Indeed, when the server is available and if no customers are present in the system, a new service (of a batch containing zero customers) might still be started with probability β. This can be interpreted as a server vacation, whose duration is characterized by the PGF $T_0(z)$.
The next step is to translate the system equations into PGFs. On account of the law of total probability, we obtain:

$$\begin{aligned} V_{k+1}(z,x,y) \triangleq& \mathrm{E}\left[z^{Q_{k+1}}x^{S_{k+1}}y^{R_{k+1}}\right] \\ =& \mathrm{E}\left[z^{Q_k+A_k}x^{S_k}y^{R_k-1}\{R_k>1\}\right] \\ &+ \mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\{R_k \leq 1, l \leq Q_k+A_k < c\}\right] \\ &+ \mathrm{E}\left[z^{Q_k+A_k-c}x^{c}y^{T_c}\{R_k \leq 1, Q_k+A_k \geq c\}\right] \\ &+ \mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\{R_k \leq 1, Q_k+A_k < l, \text{service starts}\}\right] \\ &+ \mathrm{E}\left[z^{Q_k+A_k}\{R_k \leq 1, Q_k+A_k < l, \text{service does not start}\}\right] \ , \end{aligned} \tag{2.1}$$

with

$$\mathrm{E}\left[z^X\{\text{condition}\}\right] \triangleq \mathrm{E}\left[z^X|\text{condition}\right]\Pr[\text{condition}] \quad .$$

Next, we invoke the property that the joint PGF of independent random variables equals the product of the corresponding marginal PGFs. Hence, due to the IID arrival process, the length of a new service only being dependent on the number of served customers, the probability that a new service starts if not enough customers are present being independent of all other random variables, expression (2.1) can be rewritten as:

$$\begin{aligned} V_{k+1}(z,x,y) =& \frac{A(z)}{y}\mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\{R_k > 1\}\right] \\ &+ \mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\{R_k \leq 1, l \leq Q_k+A_k < c\}\right] \\ &+ z^{-c}x^cT_c(y)\mathrm{E}\left[z^{Q_k+A_k}\{R_k \leq 1, Q_k+A_k \geq c\}\right] \\ &+ \beta\mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\{R_k \leq 1, Q_k+A_k < l\}\right] \\ &+ (1-\beta)\mathrm{E}\left[z^{Q_k+A_k}\{R_k \leq 1, Q_k+A_k < l\}\right] \quad . \end{aligned} \tag{2.2}$$

Owing to assumption 1 (the load $\rho < 1$), the queueing system under investigation eventually - i.e. for large enough values of k - reaches a steady state, implying that the distributions of all involved random variables become independent of their time index k. For instance, $V_k(z,x,y)$ and $V_{k+1}(z,x,y)$ converge to the common steady-state limit $V(z,x,y)$. Before going to the steady state, we first introduce some definitions:

$$q_0(n) \triangleq \lim_{k\to\infty}\Pr[Q_k = n, R_k = 0] \quad , \qquad 0 \leq n \leq l-1 \quad , \tag{2.3}$$

$$d(n) \triangleq \lim_{k\to\infty}\Pr[Q_k+A_k = n, R_k \leq 1] \quad , \qquad 0 \leq n \leq c-1 \quad , \tag{2.4}$$

$$F(z,x) \triangleq \lim_{k\to\infty}\mathrm{E}\left[z^{Q_k}x^{S_k}\{R_k = 1\}\right] \quad . \tag{2.5}$$

On account of the law of total probability, the IID arrival process, definitions (2.3)-(2.5) and

$$V(z,x,0) = \sum_{n=0}^{l-1} q_0(n)z^n \quad ,$$

(because $R_k = 0 \Rightarrow Q_k \leq l-1$ and $S_k = 0$), equation (2.2) evolves in the steady state to

$$\begin{aligned} V(z,x,y) =& \frac{A(z)}{y}\left\{V(z,x,y) - \sum_{n=0}^{l-1}q_0(n)z^n - yF(z,x)\right\} \\ &+ \sum_{n=l}^{c-1}d(n)x^nT_n(y) \\ &+ \left(\frac{x}{z}\right)^c T_c(y)\left[A(z)F(z,1) + A(z)\sum_{n=0}^{l-1}q_0(n)z^n - \sum_{n=0}^{c-1}d(n)z^n\right] \\ &+ \beta\sum_{n=0}^{l-1}d(n)x^nT_n(y) + (1-\beta)\sum_{n=0}^{l-1}d(n)z^n \quad . \end{aligned} \tag{2.6}$$

Note that $F(z,1)$ means that the service time is summed out from $F(z,x)$: $F(z,1) = \lim_{k\to\infty}\mathrm{E}\left[z^{Q_k}\{R_k = 1\}\right]$. Next, notice that definitions (2.3) and

(2.4) imply that

$$q_0(n) = d(n)(1-\beta) \ , \qquad 0 \leq n \leq l-1 \ . \tag{2.7}$$

Indeed, "$Q_{k+1} = n, R_{k+1} = 0$" means that the server is not processing during slot $k+1$ and that n customers are present at the beginning of that slot. This can only be the case if the server becomes available at the end of slot k ($R_k \leq 1$) and if n customers are present at that moment (i.e. $Q_k + A_k = n$) and if the server does not start service anyway at slot mark $k+1$ (with probability $(1-\beta)$). Hence, $\Pr[Q_{k+1} = n, R_{k+1} = 0] = \Pr[Q_k + A_k = n, R_k \leq 1](1-\beta)$. Since $\rho < 1$, this becomes independent of the slot index k (or $k+1$), and thus leads to expression (2.7). Substitution of (2.7) in (2.6) produces

$$\begin{aligned} V(z,x,y)\left[1-\frac{A(z)}{y}\right] =& (1-\beta)\left[1-\frac{A(z)}{y}\right]\sum_{n=0}^{l-1} d(n)z^n \\ &+ \left(\frac{x}{z}\right)^c T_c(y)[A(z)-1]\sum_{n=0}^{l-1} d(n)z^n \\ &+ \beta\sum_{n=0}^{l-1} d(n)\left[x^n T_n(y) - z^n\left(\frac{x}{z}\right)^c T_c(y)A(z)\right] \\ &+ \left(\frac{x}{z}\right)^c T_c(y)A(z)F(z,1) - A(z)F(z,x) \\ &+ \sum_{n=l}^{c-1} d(n)\left[x^n T_n(y) - z^n\left(\frac{x}{z}\right)^c T_c(y)\right] \ . \end{aligned} \tag{2.8}$$

Substituting y by $A(z)$ and letting $x \to 1$ leads to the following expression for $F(z,1)$:

$$\begin{aligned} A(z)F(z,1)\left[z^c - T_c(A(z))\right] =& T_c(A(z))[A(z)-1]\sum_{n=0}^{l-1} d(n)z^n \\ &+ \beta\sum_{n=0}^{l-1} d(n)\left[z^c T_n(A(z)) - z^n T_c(A(z))A(z)\right] \\ &+ \sum_{n=l}^{c-1} d(n)\left[z^c T_n(A(z)) - z^n T_c(A(z))\right] \ . \end{aligned} \tag{2.9}$$

Note that $T_n(A(z))$ is also a PGF and it represents the number of arriving customers during a service of a batch of n customers. Next, substituting y by

$A(z)$ in (2.8) and appealing to (2.9) yields a formula for $F(z,x)$:

$$\begin{aligned}
&z^c A(z)F(z,x)\left[z^c - T_c(A(z))\right]\\
=&z^c x^c T_c(A(z))[A(z)-1]\sum_{n=0}^{l-1} d(n)z^n\\
&+\beta x^c T_c(A(z))\sum_{n=0}^{l-1} d(n)\left[z^c T_n(A(z)) - z^n T_c(A(z))A(z)\right]\\
&+x^c T_c(A(z))\sum_{n=l}^{c-1} d(n)\left[z^c T_n(A(z)) - z^n T_c(A(z))\right]\\
&+\beta[z^c - T_c(A(z))]\sum_{n=0}^{l-1} d(n)\left[z^c x^n T_n(A(z)) - x^c z^n T_c(A(z))A(z)\right]\\
&+[z^c - T_c(A(z))]\sum_{n=l}^{c-1} d(n)\left[z^c x^n T_n(A(z)) - x^c z^n T_c(A(z))\right] \ . \qquad (2.10)
\end{aligned}$$

Remark 2. *When $\beta = 0$ and $T_n(z) = T_c(z)$, $\forall n$, expression (2.8) transforms, by relying on (2.9) and (2.10), into formula (4) from our paper [43]. In [43], we have studied several aspects, including the joint PGF of the queue content, the server content and the remaining service time, of a model that adopts the threshold-based service policy and whereby the service times are independent of the number of customers in the served batches.*

Expressions (2.8)-(2.10) provide enough information to deduce a sprectrum of quantities related to the buffer content, which forms the subject of the next section. However, formulas (2.8)-(2.10) still contain the unknown probabilities $d(n)$. In order to explain how these can be calculated, set $y = 1$ and $x = z$ in (2.8), leading to

$$V(z,z,1)[1-A(z)] = A(z)F(z,1) - A(z)F(z,z) \ ,$$

which can, by applying (2.9) and (2.10), be transformed into

$$\begin{aligned}
V(z,z,1)[1-A(z)]\left[z^c - T_c(A(z))\right] =&(z^c-1)T_c(A(z))[1-A(z)]\sum_{n=0}^{l-1} d(n)z^n\\
&+\beta\sum_{n=0}^{l-1} d(n)g_n(z) + \sum_{n=l}^{c-1} d(n)h_n(z) \ , \qquad (2.11)
\end{aligned}$$

with

$$\begin{aligned}
g_n(z) \triangleq&(z^n - z^c)T_n(A(z))T_c(A(z)) + z^n(z^c-1)T_c(A(z))A(z)\\
&- z^c(z^n-1)T_n(A(z)) \ , \qquad (2.12)\\
h_n(z) \triangleq&T_n(A(z))z^c\{1 - z^n - T_c(A(z))\} - T_c(A(z))z^n\{1 - z^c - T_n(A(z))\} \ . \qquad (2.13)
\end{aligned}$$

One can prove by means of Rouché's theorem that $z^c - T_c(A(z))$ has c zeroes $(z_0, z_1, \ldots, z_{c-1})$ in the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$ (see e.g. [5]). On account of the normalization condition of PGFs ($T_c(A(1)) = 1$), we find that one of these zeroes, say z_0, equals 1. The other zeroes z_i can be calculated one-by-one, by means of a standard root-finding algorithm such as Newton-Raphson, by solving the $c-1$ equations

$$z_i = T_c(A(z_i))^{1/c}\varepsilon_i \ , \qquad 1 \leq i \leq c-1 \ ,$$

with $\varepsilon_i \triangleq e^{\imath 2i\pi/c}$ and $\imath$ the imaginary unit[1] (the ε_i's, together with $\varepsilon_0 = 1$, thus represent the c-th roots of 1). Because $V(z, z, 1)$ is a PGF and since PGFs are normalized ($V(1, 1, 1) = 1$) and bounded (i.e. they have no poles) in the closed complex unit disk, the unknowns $d(n)$ can be determined by solving a set of c linear equations[2], consisting of the normalization condition and $c-1$ equations expressing that the numerator of $V(z, z, 1)$ (i.e. the right-hand-side of (2.11)) vanishes at z_i, $1 \leq i \leq c-1$ (for $i = 0$, it leads to the trivial equation $0 = 0$, which thus produces no information):

$$\begin{aligned}[1 - A(z_i)] \sum_{n=0}^{l-1} d(n) z_i^n + \beta \sum_{n=0}^{l-1} d(n) \left[A(z_i) z_i^n - T_n(A(z_i))\right] \\ + \sum_{n=l}^{c-1} d(n)[z_i^n - T_n(A(z_i))] = 0 \ , \qquad 1 \leq i \leq c-1 \ , \end{aligned} \tag{2.14}$$

$$\begin{aligned}-c + \mathrm{E}[T_c]\,\lambda = - c \sum_{n=0}^{l-1} d(n) + \beta \sum_{n=0}^{l-1} d(n)[c + n\mathrm{E}[T_c] - c\mathrm{E}[T_n]] \\ + \sum_{n=l}^{c-1} d(n)[n\mathrm{E}[T_c] - c\mathrm{E}[T_n]] \ , \end{aligned} \tag{2.15}$$

whereby we have also taken into account that $T_c(A(z_i)) = z_i^c$ in (2.14). Owing to the one-on-one relation between a discrete probability distribution and its PGF and the uniqueness of the stationary distribution of the system content at slot marks (because the stability condition holds: assumption 1), this set of equations has a unique solution. Now, we have all the tools at our disposal required to deduce all kinds of quantities related to the buffer content.

Remark 3. *Equation (2.15) can also be established by utilizing that in a system in steady state the average number of customers that enter the system in a slot (λ) equals the average number of customers that leave the system in a slot:*

$$\begin{aligned}\lambda &= \lim_{k\to\infty} \mathrm{E}[S_k\{R_k = 1\}] \\ &= \frac{d}{dx} F(1, x)\bigg|_{x=1} \ . \end{aligned}$$

2.3 Quantities related to the buffer content

In this section, we extract from formulas (2.8)-(2.10) a wide spectrum of quantities related to the buffer content. Let us start with the system content at random slot marks.

[1] Throughout this dissertation, $z^{1/c}$ represents the principal branch of the complex c-th root function, i.e. $z^{1/c} \triangleq |z|^{1/c} e^{\imath \mathrm{Arg}(z)/c}$ with $\mathrm{Arg}(z)$ the principal value of the argument of z, i.e. a mapping in the interval $] - \pi, \pi]$.

[2] If $z^c - T_c(A(z))$ is periodic, one cannot always determine the unknowns by solving a set of c linear equations (for instance when $c = 2k$, $l = c$, $\beta = 0$ and $A(z) = \sum_{n=0}^{\infty} a(2n) z^{2n}$). The reason stems from the fact that when the period equals p, p zeroes from the set of zeroes z_i are also zeroes of $z^p - 1$ and thus of $z^c - 1$ (this is proved in [5]). Instead, one should then use ad hoc arguments to reduce the problem into a solvable model (we refer to our paper [43] for an example).

2.3.1 System content at random slot boundaries

As the system content U equals the sum of the queue and the server content, its PGF $U(z)$ is, owing to the properties of joint PGFs, equal to $V(z,z,1)$. Hence, substituting $V(z,z,1)$ by $U(z)$ in expression (2.11) yields

$$\begin{aligned} U(z)[1-A(z)]\,[z^c - T_c(A(z))] =& (z^c-1)T_c(A(z))[1-A(z)]\sum_{n=0}^{l-1} d(n)z^n \\ &+ \beta \sum_{n=0}^{l-1} d(n)g_n(z) + \sum_{n=l}^{c-1} d(n)h_n(z)\ , \end{aligned} \tag{2.16}$$

whereby $g_n(z)$ and $h_n(z)$ are defined by respectively (2.12) and (2.13). In the special case $T_n(z) = z$, $\forall n$, $l = 1$ and $\beta = 0$, (2.16) transforms into expression (4) from [117], where the PGF of the system content at random slot marks in the $Geo^X/D^{1,c}/1$ queueing model is obtained. In addition, when $\beta = 0$, we find formula (13) from our paper [38] and when $\beta = 0$ and $T_n(z) = T_c(z)$, $\forall n$, we obtain expression (9) from [37].

2.3.2 Queue content at random slot boundaries

Next, we continue with the queue content at random slot marks. Its PGF $Q(z)$ is found by summing out both the server content and the remaining service time from the joint PGF $V(z,x,y)$. Hence, letting $y \to 1$ and $x \to 1$ in (2.8) and applying (2.9), we find

$$\begin{aligned} &Q(z)[1-A(z)]\,[z^c - T_c(A(z))] \\ =&(z^c-1)[1-A(z)]\sum_{n=0}^{l-1} d(n)z^n \\ &+\beta\sum_{n=0}^{l-1} d(n)\,[(1-z^n)\{z^c - T_c(A(z))\} + (z^c-1)\{z^n A(z) - T_n(A(z))\}] \\ &+\sum_{n=l}^{c-1} d(n)\,[z^c - z^n + (z^n-1)T_c(A(z)) + (1-z^c)T_n(A(z))]\ . \end{aligned} \tag{2.17}$$

Note that $U(z) \neq T_c(A(z))Q(z)$ as in [74]. In [74], it is shown that for a broad class of discrete- and continuous-time queueing systems, the stationary system content is the sum of two independent random variables, one of which is the stationary queue content and the other is the number of customers that arrive during the time a customer spends in service. The relation $U(z) = T_c(A(z))Q(z)$ does not hold here because the service time of a batch is not independent of the number of customers within it, which is one of the necessary conditions in [74]. When $T_n(z) = T_c(z)$, $\forall n$, then we indeed find that $U(z) = T_c(A(z))Q(z)$.

2.3.3 System content at service completion times

The system content $\tilde{U}$ at service completion times[3] equals the sum of the queue content at the beginning of the final slot of the service and the number of customers that have arrived during that slot. Hence, from the definition of $F(z,x)$ in (2.5), we get

$$\tilde{U}(z) = A(z)\frac{F(z,1)}{F(1,1)} \ ,$$

(note that the division by $F(1,1)$ is necessary to assure that $\tilde{U}(1) = 1$). On account of expression (2.9) for $F(z,1)$, this can be rewritten as

$$\begin{aligned}\tilde{U}(z) = \frac{1}{F(1,1)[z^c - T_c(A(z))]} & \Bigg[T_c(A(z))[A(z)-1] \sum_{n=0}^{l-1} d(n) z^n \\ & + \beta \sum_{n=0}^{l-1} d(n) \left\{ z^c T_n(A(z)) - z^n T_c(A(z))A(z) \right\} \\ & + \sum_{n=l}^{c-1} d(n) \left\{ z^c T_n(A(z)) - z^n T_c(A(z)) \right\} \Bigg] \ , \qquad (2.18)\end{aligned}$$

with

$$\begin{aligned}F(1,1) = \frac{1}{c - \mathrm{E}[T_c]\lambda} & \Bigg[\lambda \sum_{n=0}^{l-1} d(n) + \beta \sum_{n=0}^{l-1} d(n) \left\{ c + \mathrm{E}[T_n]\lambda - n - \mathrm{E}[T_c]\lambda - \lambda \right\} \\ & + \sum_{n=l}^{c-1} d(n) \left\{ c + \mathrm{E}[T_n]\lambda - n - \mathrm{E}[T_c]\lambda \right\} \Bigg] \ . \qquad (2.19)\end{aligned}$$

When $l = 0$, $\beta = 0$, $A(z) = 1 - \lambda + \lambda z$ and $T_n(z) = T_c(z)$ in (2.18), we find formula (4.4.1) from [32]. In [32], the system content at service completion times has been studied for the following queueing models with batch service: the continuous time $M/G^{(0,c)}/1$, $M/G^{(0,c)}/1/K$ and $M/G^{(c,c)}/1$ queues and the discrete time $Geom/G^B/1$ queue. Note that, as opposed to the model in this dissertation, the models in [32] do not include batch arrivals, a service threshold, a timer mechanism and service times being dependent on the number of customers within it.

2.3.4 Server content at random slot boundaries

The preceding quantities are in particular of importance from a customer point of view. Indeed, the larger the queue or system content, the longer the delay that customers suffer. In the following, we establish quantities that are especially useful from an economic point of view. In practice, service might be expensive and therefore it is desired to exploit service capacity efficiently. We now establish quantities that describe the efficiency of the server usage. Let us start with the server content at random slot marks.

[3] Note that we again observe immediately after the potential start of service epoch, but now conditioned on the event that a service just has terminated.

Its PGF $S(z)$ is found by summing out both the queue content and the remaining service time from $V(z,x,y)$. Hence, by letting $y \to 1$ and $z \to 1$ in (2.8), thereby applying l'Hôpital's rule and relying on (2.10) and finally substituting x by z, we obtain

$$\begin{aligned}
&S(z)\left[c - \mathrm{E}\left[T_c\right]\lambda\right] \\
=&(1-\beta)\left[c - \mathrm{E}\left[T_c\right]\lambda\right]\sum_{n=0}^{l-1} d(n) + \beta\left[c - \mathrm{E}\left[T_c\right]\lambda\right]\sum_{n=0}^{l-1} d(n)z^n\mathrm{E}\left[T_n\right] \\
&+ \left[c - \mathrm{E}\left[T_c\right]\lambda\right]\sum_{n=l}^{c-1} d(n)z^n\mathrm{E}\left[T_n\right] + z^c\lambda\mathrm{E}\left[T_c\right]\sum_{n=0}^{l-1} d(n) \\
&+ z^c\mathrm{E}\left[T_c\right]\beta\sum_{n=0}^{l-1} d(n)\{\mathrm{E}\left[T_n\right]\lambda - n - \lambda\} + z^c\mathrm{E}\left[T_c\right]\sum_{n=l}^{c-1} d(n)\{\mathrm{E}\left[T_n\right]\lambda - n\} \ . \qquad (2.20)
\end{aligned}$$

Note that $S(z)$ is a polynomial of degree c, which allows us, owing to the definition of a PGF, to easily extract the corresponding probabilities (we substitute $(1-\beta)d(n)$ by $q_0(n)$):

$$\begin{aligned}
&\Pr\left[S = n\right] = \\
&\begin{cases}
\sum_{m=0}^{l-1} q_0(m) + \beta\mathrm{E}\left[T_0\right]d(0) & \text{if } n = 0 \ , \\
\beta\mathrm{E}\left[T_n\right]d(n) & \text{if } 1 \le n \le l-1 \ , \\
\mathrm{E}\left[T_n\right]d(n) & \text{if } l \le n \le c-1 \ , \\
1 - \sum_{m=0}^{l-1} q_0(m) - \beta\sum_{m=0}^{l-1}\mathrm{E}\left[T_m\right]d(m) - \sum_{m=l}^{c-1}\mathrm{E}\left[T_m\right]d(m) & \text{if } n = c \ , \\
0 & \text{else } .
\end{cases}
\end{aligned}$$

2.3.5 Number of customers in a served batch

The number of customers in a random served batch, $\tilde{S}$, is, because newly arriving customers cannot join the ongoing service, equally distributed as the server content at the last slot of the service period, which yields

$$\tilde{S}(z) = \frac{F(1,z)}{F(1,1)} \ ,$$

and in view of expression (2.10) for $F(z,x)$, we find, after application of l'Hôpital's rule,

$$\begin{aligned}
\tilde{S}(z) = \frac{1}{F(1,1)[c - \mathrm{E}\left[T_c\right]\lambda]}&\left[z^c\lambda\sum_{n=0}^{l-1} d(n) + \beta z^c\sum_{n=0}^{l-1} d(n)\left\{\mathrm{E}\left[T_n\right]\lambda - n - \lambda\right\}\right. \\
&+ z^c\sum_{n=l}^{c-1} d(n)\left\{\mathrm{E}\left[T_n\right]\lambda - n\right\} + \beta\left\{c - \mathrm{E}\left[T_c\right]\lambda\right\}\sum_{n=0}^{l-1} d(n)z^n \\
&\left. + \left\{c - \mathrm{E}\left[T_c\right]\lambda\right\}\sum_{n=l}^{c-1} d(n)z^n\right] \ . \qquad (2.21)
\end{aligned}$$

Notice that $\tilde{S}(z)$ is a polynomial of degree c, which allows us to easily extract the corresponding probabilities:

$$\Pr\left[\tilde{S} = n\right] = \begin{cases}
\frac{\beta d(n)}{F(1,1)} & \text{if } 0 \le n \le l-1 \ , \\
\frac{d(n)}{F(1,1)} & \text{if } l \le n \le c-1 \ , \\
1 - \frac{\beta\sum_{m=0}^{l-1} d(m) + \sum_{m=l}^{c-1} d(m)}{F(1,1)} & \text{if } n = c \ , \\
0 & \text{else } .
\end{cases}$$

Remark 4. *From the probability distribution* $\Pr\left[\tilde{S}=n\right]$ *it is possible to calculate the service time of a randomly tagged customer. By taking into account that a randomly tagged customer is more likely to be served in a batch with more customers (see e.g. [24]), the probability* $\Pr\left[\hat{S}=n\right]$ *that this customer is served in a batch with* n *customers equals*

$$\Pr\left[\hat{S}=n\right]=\frac{\Pr\left[\tilde{S}=n\right]n}{\mathrm{E}\left[\tilde{S}\right]} \quad .$$

As a result, the PGF of the service time of a randomly tagged customer reads

$$\sum_{n=0}^{c}\frac{\Pr\left[\tilde{S}=n\right]n}{\mathrm{E}\left[\tilde{S}\right]}T_n(z) \quad .$$

2.3.6 Probability that the server processes

The probability that the server processes a batch during a random slot ensues almost immediately from the fact that the server is not serving if and only if $R_k=0$, the observation that $R_k=0 \Rightarrow 0 \le Q_k \le l-1$, the law of total probability and the definition of $q_0(n)$ in (2.3):

$$\Pr\left[\text{server processes}\right]=1-\sum_{n=0}^{l-1}q_0(n) \quad . \tag{2.22}$$

2.3.7 Queue content when the server not processes

The final quantity we deduce is the queue content at a random slot mark given that the server is not serving. Its PGF $\tilde{Q}(z)$ is found by taking into account that the server is not processing if and only if the remaining service time equals 0. The server content being zero is a necessary but not a sufficient condition because the server can process a batch containing zero customers. Hence,

$$\tilde{Q}(z)=\frac{V(z,0,0)}{V(1,0,0)}=\frac{\sum_{n=0}^{l-1}q_0(n)z^n}{\sum_{m=0}^{l-1}q_0(m)} \quad . \tag{2.23}$$

The corresponding probabilities thus are:

$$\Pr\left[\tilde{Q}=n\right]=\begin{cases}\frac{q_0(n)}{\sum_{m=0}^{l-1}q_0(m)} & \text{if } 0\le n\le l-1 \ , \\ 0 & \text{else } .\end{cases}$$

This quantity is of importance as it gives an indication of the number of customers suffering an extra delay because service is postponed (postponing delay).

Remark 5. *Note that when* $l=0$ *or* $\beta=1$*, it makes no sense to calculate* $\Pr\left[\tilde{Q}=n\right]$*, as the server always processes in that case.*

2.4 Performance measures

The performance of actual batch-service systems is often assessed by evaluating various moments of the abovementioned quantities (2.16)-(2.23). As we have assumed that the radius of convergence of each PGF is larger than one (assumption 2), all order moments exist and can be calculated by applying the moment generating property of PGFs. In addition, as buffers have a finite capacity in practice, the loss ratio - defined as the fraction of customers that cannot enter the system due to a full queue - has to be assessed. Next, one often has to dimension the queue size so that the loss ratio is below some threshold. Although we assume an infinite buffer capacity in this dissertation, it has been shown in [26] that $\Pr[Q > b]$ for an infinite buffer system provides a good approximation for the loss ratio of the corresponding system with buffer capacity b, when b is large (which is the case in real life, because the loss-ratio has to be kept small). However, for large b, it becomes unfeasible to calculate $\Pr[Q > b]$ via the probability generating property of PGFs, as it would require b-th order derivatives to be taken. Therefore, we resort to the approximation technique mentioned in section 1.4. First, on account of assumptions 1-5, one can prove, completely analogously as in [106], that $Q(z)$ has one dominant singularity, say $\tilde{z}$, that it is a pole with multiplicity one and that it is the only zero of $z^c - T_c(A(z))$ in $]1, \Re[$[4]. Consequently, application of formula (1.4) yields

$$\Pr[Q > b] \approx \frac{\tilde{z}^{-(b+1)}}{1-\tilde{z}} \frac{N_Q(\tilde{z})}{D_Q^{'}(\tilde{z})} \ , \tag{2.24}$$

with $N_Q(z)$ and $D_Q(z)$ respectively the numerator and denominator of $Q(z)$. Consequently, the minimum buffer capacity b required to assure that $\Pr[Q > b] < 10^{-m}$ (m is some integer) is found by taking the Neperian logarithm of this equation and on account of (2.24), we find

$$b = \left\lceil \frac{\ln\left(\frac{N_Q(\tilde{z})}{(1-\tilde{z})D_Q^{'}(\tilde{z})}\right) + m\ln(10)}{\ln(\tilde{z})} \right\rceil - 1 \ .$$

2.5 Numerical examples

In this section, we demonstrate that the abovementioned performance measures are useful tools to evaluate batch-service queueing systems. We consider an example whereby

- The number of customer arrivals during a slot is Poisson distributed (i.e. $A(z) = e^{\lambda(z-1)}$).
- The server capacity c is equal to 10.
- The service times are geometrically distributed with mean length being dependent on the number of customers (say n) in the served batch: $\mathrm{E}[T_n] = 8 + 0.2n$. The mean service time thus consists of a constant part and a part dependent on n. Notice that we have opted for a long constant part as we feel this is more realistic. For instance, in telecommunications, the construction of the header for the batch typically takes a longer time than the actual transmission. Also, delivering goods has a

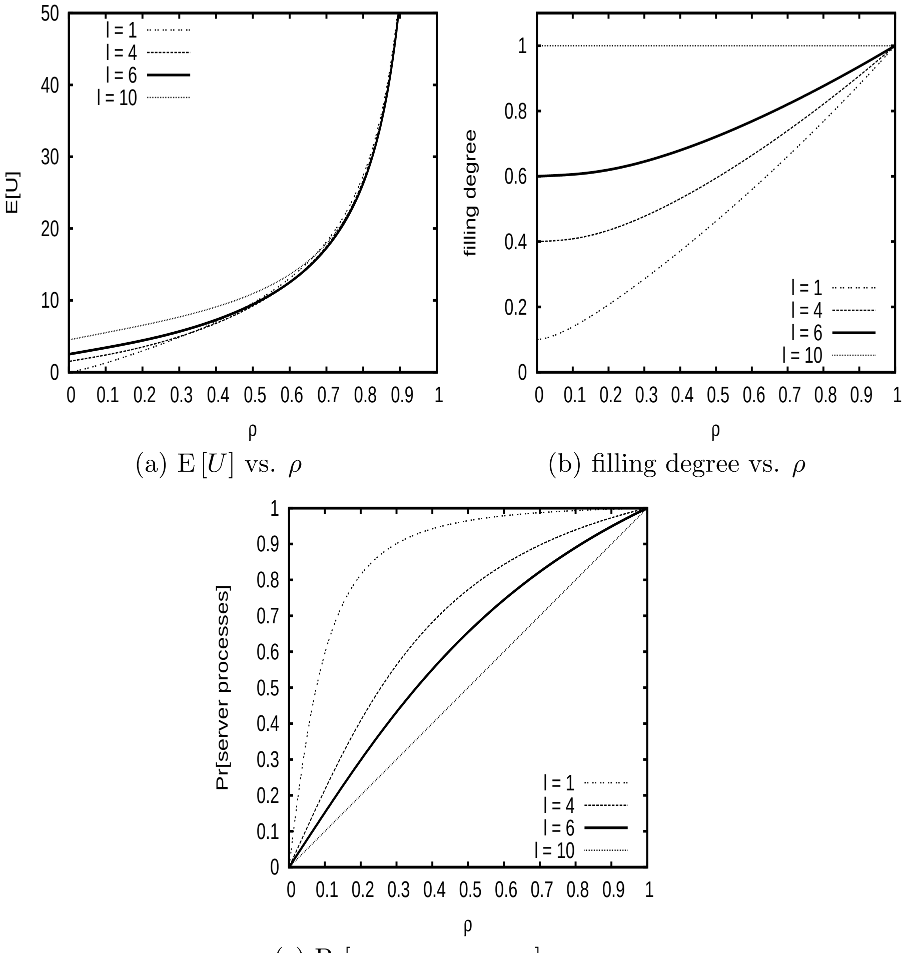
[4] The zero $\tilde{z}$ can be calculated by means of a standard root-finding algorithm such as Newton-Raphson or, because $\tilde{z}$ lies on the real axis, via the bisection method.

long constant part (e.g. driving on the highway) and a smaller dependent part (e.g. loading and unloading).

First, we evaluate the influence of the service threshold l when $\beta = 0$. We have therefore depicted the mean system content, the filling degree (i.e. the mean number of customers in a served batch divided by the server capacity) and the probability that the server processes a batch in a random slot versus the load ρ in Fig. 2.1 (recall that the load was defined in chapter 1 as $\rho \triangleq \lambda \mathrm{E}\left[T_c\right]/c$). We notice that in case of low load (also called light traffic), a threshold larger than one leads to a much larger mean system content when no mechanism exists to avoid long postponing delays (i.e. $\beta = 0$). In case of higher loads, larger thresholds become preferable, because when more customers arrive, it pays off to wait in order to exploit the server capacity. When the load tends to one (also called heavy traffic), the influence of the thresholds fades, as nearly always c customers will be served anyway. The figures also clearly demonstrate that the larger the threshold, the better the server capacity is utilised (i.e. a larger filling degree and a smaller probability that the server processes).

Next, we investigate the influence of the probability β to start a service even when less than l customers are present. Therefore, in Fig. 2.2, the mean system content is depicted versus the load for $l = 5$ and several values of β (part a) and versus β for $\rho = 0.01$ (part b) and $\rho = 0.7$ (part c).
The figure exhibits that in case of light traffic $\beta = 1$ is the best option, whereas it is preferable to select a smaller value for β when the load becomes larger. Indeed, when customers arrive seldomly, it is better not to wait, whereas it pays off to postpone service when customers arrive frequently. The figures also demonstrate that $\beta = 0.1$ always leads to reasonable results and thus is a good selection if the load is not known a priori.

Before closing this section, we investigate the influence of β and l on the queue capacity that is required to assure that the loss ratio is smaller than 10^{-6} (Fig. 2.3). We perceive that in case of light traffic $l > 1$ and $\beta = 0$ leads to some larger required buffer capacity. The reason is that customers stay in the queue, whereas when $\beta \neq 0$ or $l = 1$, customers leave the system almost immediately. When the load becomes larger, the required buffer capacity increases and the influence of l and β fades.

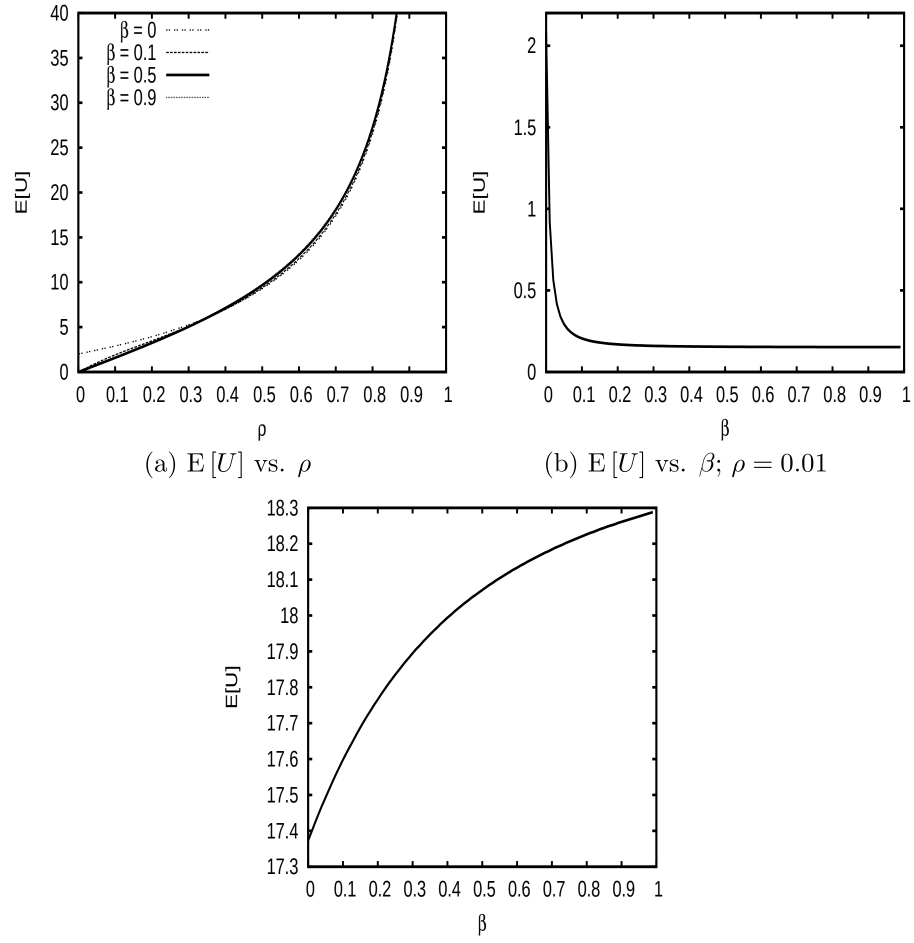


(a) $\mathrm{E}[U]$ vs. ρ

(b) filling degree vs. ρ

(c) Pr[server processes] vs. ρ

Figure 2.1: Influence of service threshold l on the behaviour of the system; Poisson arrivals, $\beta = 0$, $c = 10$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$

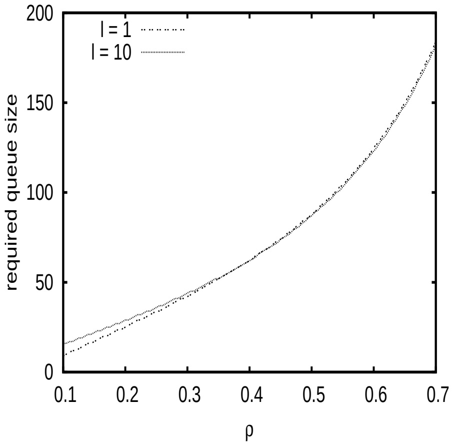


(a) $\mathrm{E}\left[U\right]$ vs. ρ

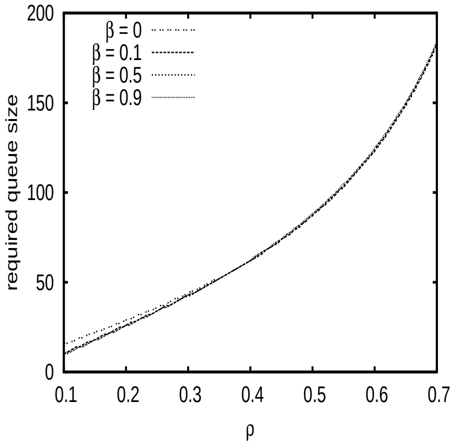
(b) $\mathrm{E}\left[U\right]$ vs. β; $\rho = 0.01$

(c) $\mathrm{E}\left[U\right]$ vs. β; $\rho = 0.7$

Figure 2.2: Influence of β on the behaviour of the system; Poisson arrivals, $c = 10$, $l = 5$, T_n geometrically distributed, $\mathrm{E}\left[T_n\right] = 8 + 0.2n$



(a) several values of l; $\beta = 0$



(b) several values of β; $l = 5$

Figure 2.3: Required queue size to assure that the loss ratio is smaller than 10^{-6} versus the load; Poisson arrivals, $c = 10$, T_n geometrically distributed, $\mathrm{E}\,[T_n] = 8 + 0.2n$

2.6 Application: group testing

In this section, we discuss an application for which our results are useful: group testing. The discussion in this section is mainly based on our paper [42].

2.6.1 Background

Classification of items as good or bad occurs in a wide area of applications. Often these items are group testable. This means that they can be tested in groups, so that, when a group test returns good, it can be concluded that all items within it are good, whereas the opposite result implies that the group contains at least one bad item. In some applications the whole group is thrown away in the latter case. Hence, no retesting is required in this so-called *incomplete-identification* scenario ([13]; [14]; [15]; [16]). When, on the other hand, the bad items need to be separated from the good, i.e. complete identification is necessary, retesting is required. This can, for instance, be achieved by testing all items of the group individually. This is often referred to as *group-individual* screening ([2]). However, often the group is divided into subgroups which are each subjected to a new group test. When adopting this *group-subgroup* strategy, one also has to choose the number of subgroups and their respective sizes. In order to avoid confusion, we adopt the term *group screening* for the complete process, i.e. for the first test on the entire (original) group and the possibly other tests on subgroups or individual items of the group. The *group size* refers to the number of items in a group or subgroup and the *original group size* is the number of items making up the original group.
Dorfman [49] was the first to introduce the paradigm of group screening and he found an immediate application in the detection of syphilitic men drafted into military service during WWII. He suggested to apply this procedure also to manufacturing processes where the defective goods have to be eliminated from the collection of all produced goods. Dorfman concluded that when a complete elimination of defective items is desired, significant savings in effort and expense can be obtained by group screening, if the prevalence rate is low and the original group size is chosen properly. Later on, several researchers applied this paradigm to HIV screening practice. For instance, Emmanuel et al. [53] performed a case study of HIV testing in Zimbabwe. They concluded that pooling five samples for HIV screening may result in a substantial reduction in costs. In addition, they mentioned that in countries where the prevalence of HIV is higher than the 23% found in Zimbabwan donors, savings may not be as great. Behets et al. [17] came to similar conclusions in a case study in Zaire. Other papers about this application include [2], [109] and [113]. The range of application even goes further. Macula [82], [83] applied this method to DNA library screening and Xie et al. [114] and Zhu et al. [118] utilized it for drug discovery. Furthermore, Sobel and Groll [105] were motivated by another practical need, this time from the industrial sector, to remove all leakers from a set of devices. In this case, one chemical apparatus is available and the

devices are tested by putting several of them in a bell jar and testing whether any of the gas used in constructing the devices has leaked out into the bell jar. Finally, in the monograph Du and Hwang [52], it is stated that group testing also emerged from many nontesting situations, such as experimental designs, multiaccess communication, coding theory, clone library screening, nonlinear optimization, and computational complexity.

Optimization of the original group size in terms of the minimization (or maximization) of some variable has been a popular research topic. However, Abolnikov and Dukhovny [2] correctly remarked that only few authors have taken into account the *dynamic nature of the item arrivals*. A dynamic nature is proper to many practical situations: items arrive at the testing center in groups of different and random size, at random moments in time. This entails that less items than the optimal original group size might be present at the time a new group can be screened. Hence, an extra decision has to be made: "When is it allowed to start screening a group with less samples than the original group size?". Abolnikov and Dukhovny [2] state that the dynamic nature can be dealt with by applying methods of queueing theory. Since then, Bar-Lev et al. [16] also made use of a queueing model to include the dynamic item arrivals.

As opposed to [2] and [16], the model in this dissertation includes a timer mechanism and we have deduced a spectrum of quantities instead of only the system content at the end of services.

2.6.2 Group-screening policies

As the model considered in this dissertation includes a general dependency between the service time of a batch and the number of items within it, a whole range of screening policies can be studied by defining the PGFs $T_j(z)$ appropriately. In this section, we demonstrate this for several group-screening policies. First, we recall that several tests might be necessary to screen a group. The number of tests is of course dependent on the service policy and the number of items in the group. Let us introduce $\tilde{T}_j$ as the number of tests required to screen a group of j items. Further, we take into account that the time to execute one test can vary, and we assume that the testing times are independent of the number of items and that they are IID with common PGF $V(z)$ (the case whereby the testing times are dependent on the number of items can also be included by defining V_j as the testing time of a group of j items). Hence, $T_j(z) = \tilde{T}_j(V(z))$. In other words, in order to study some screening policy, one has to define $\tilde{T}_j(z)$ properly. We illustrate this by considering incomplete identification, the group-individual testing and a group-subgroup screening policy. In the latter policy, when the outcome of a test on a group of j items is bad, two subgroups of $\lceil j/2 \rceil$ and $\lfloor j/2 \rfloor$[5] items are subjected to new group tests and after a bad subgroup test, all items from the bad subgroup are

[5] We adopt the standard convention that $\lceil . \rceil$ and $\lfloor . \rfloor$ represent respectively the ceil and the floor function, i.e. $\lceil x \rceil \triangleq \min\{n \in \mathbb{Z} : n \geq x\}$ and $\lfloor x \rfloor \triangleq \max\{n \in \mathbb{Z} : n \leq x\}$.

retested individually. Before studying the policies, it is important to point out that only one test is required when $j = 1$, so that

$$\tilde{T}_1(z) = z \ .$$

In the remainder, we designate the probability of a random item being bad by p (and let $\overline{p} \triangleq 1 - p$) and we assume that this probability is independent of the result of the other items.

Incomplete identification

When the outcome of the group test is good, all items within the group are classified as good, whereas, in the opposite case, the group is 'thrown away'. Either way, only one test is required, leading to

$$\tilde{T}_j(z) = z \ , \qquad j \geq 2 \ .$$

Group-individual testing

Two situations occur here. In the first case, none of the items is bad (with probability $\overline{p}^j$), which leads to a good test result, implying that no retesting is required. In the second case, one or more items are bad (with probability $1 - \overline{p}^j$), so that the j items need to be retested, leading to $j + 1$ tests. It can be concluded that:

$$\tilde{T}_j(z) = \overline{p}^j z + (1 - \overline{p}^j) z^{j+1} \ , \qquad j \geq 2 \ .$$

A group-subgroup testing policy

If the outcome is bad in this scenario, the group is split in two subgroups which are both retested. If the outcome of a test on a subgroup is bad, then all the items within it are retested individually. In order to construct $\tilde{T}_j(z)$, we make a distinction between $j = 2$, $j = 3$ and $j \geq 4$.

$j = 2$

If the group test returns bad, two subgroups of 1 item require each only one extra test. Hence,

$$\tilde{T}_2(z) = \overline{p}^2 z + (1 - \overline{p}^2) z^3 \ .$$

$j = 3$

After a bad result, a subgroup of one and a subgroup of two items are retested. The subgroup of size two needs another group test, and if this returns bad again, two extra tests are required. Hence,

$$\tilde{T}_3(z) = \overline{p}^3 z + \overline{p}^2 p z^3 + (1 - \overline{p}^2) z^5 \ .$$

$j \geq 4$

Let us assume that the first subgroup contains $\lceil j/2 \rceil$ and the second $\lfloor j/2 \rfloor$ items (when a retest is necessary). This case is split in four subcases, according to the number of bad items, i:

- $i = 0$ (with probability $\overline{p}^j$); only 1 test is required.
- $i = 1$. Two subcases:
  - The bad item belongs to the first part (with probability $\lceil j/2 \rceil\, p\overline{p}^{j-1}$); $\lceil j/2 \rceil + 3$ tests are required, namely the original test, the two subgroup tests and $\lceil j/2 \rceil$ individual tests.
  - The bad item belongs to the second part (with probability $\lfloor j/2 \rfloor\, p$ $\overline{p}^{j-1}$); $\lfloor j/2 \rfloor + 3$ tests are required.
- $2 \leq i \leq \lceil j/2 \rceil$. Three subcases arise:
  - All the bad items belong to the first subgroup (with probability $\binom{\lceil j/2 \rceil}{i} p^i \overline{p}^{j-i}$); $\lceil j/2 \rceil + 3$ tests are required.
  - All the bad items belong to the second subgroup (with probability $\binom{\lfloor j/2 \rfloor}{i} p^i \overline{p}^{j-i}$); $\lfloor j/2 \rfloor + 3$ tests are required. Note that this case cannot occur when $i = \lceil j/2 \rceil$ and j odd.
  - The bad items are spread over both parts of the group (with probability $\left[\binom{j}{i} - \binom{\lceil j/2 \rceil}{i} - \binom{\lfloor j/2 \rfloor}{i}\right] p^i \overline{p}^{j-i}$); $j + 3$ tests are required.
- $\lceil j/2 \rceil < i \leq j$ (with probability $\binom{j}{i} p^i \overline{p}^{j-i}$); $j + 3$ tests are required, as the bad items are certainly spread over both subgroups.

Summarized, $\tilde{T}_j(z)$ equals:

$$\tilde{T}_1(z) = z \ , \tag{2.25}$$

$$\tilde{T}_2(z) = \overline{p}^2 z + (1 - \overline{p}^2) z^3 \ , \tag{2.26}$$

$$\tilde{T}_3(z) = \overline{p}^3 z + \overline{p}^2 p z^3 + (1 - \overline{p}^2) z^5 \ , \tag{2.27}$$

$$\begin{aligned}
\tilde{T}_j(z) =& \overline{p}^j z \\
&+ \left[\lfloor j/2 \rfloor\, p\overline{p}^{j-1} + \sum_{i=2}^{\lceil j/2 \rceil} \binom{\lfloor j/2 \rfloor}{i} p^i \overline{p}^{j-i} \right] z^{\lfloor j/2 \rfloor + 3} \\
&+ \left[\lceil j/2 \rceil\, p\overline{p}^{j-1} + \sum_{i=2}^{\lceil j/2 \rceil} \binom{\lceil j/2 \rceil}{i} p^i \overline{p}^{j-i} \right] z^{\lceil j/2 \rceil + 3} \\
&+ \left[\sum_{i=2}^{\lceil j/2 \rceil} \left\{ \binom{j}{i} - \binom{\lceil j/2 \rceil}{i} - \binom{\lfloor j/2 \rfloor}{i} \right\} p^i \overline{p}^{j-i} \right. \\
&+ \left. \sum_{i=\lceil j/2 \rceil + 1}^{j} \binom{j}{i} p^i \overline{p}^{j-i} \right] z^{j+3} \ , \qquad j \geq 4 \ .
\end{aligned} \tag{2.28}$$

Remark 6. *The model, in fact, also includes the individual screening policy; one just has to set $\tilde{T}_j(z)$ to be equal to z^j.*

2.6.3 Example

Evaluating a practical group-testing scenario boils down to plug in the right parameters in the queueing model, to calculate performance measures for various values of l, β and c and select those values that lead to the best results. In order to illustrate this, consider a blood testing centre whereby a laboratory assistant checks every 5 minutes if a new group can be screened. This is the case when the test kit is available and when the number of available blood samples reaches or exceeds l. Next, new samples arrive according to a Poisson process with parameter 0.05 (samples/minute). Further, it takes 50 minutes to complete one test and a random blood sample is infected with probability 0.025. Finally, the group-subgroup screening procedure as in section 2.6.2 is adopted.

This setting can be fit in our model by making the following assumptions:

- Slots correspond to periods of 5 minutes and the slot marks correspond to the check times.
- Since the arrival process is a Poisson process, the number of arrivals in a slot is Poisson distributed with mean equal to 0.05 × 5, i.e., $A(z) = e^{0.25(z-1)}$.
- $\beta = 0$.
- $p = 0.025$.
- A test takes 10 slots, implying that $V(z)$ is equal to z^{10}; hence, $T_j(z) = \tilde{T}_j(z^{10})$.
- $\tilde{T}_j(z)$ obeys formulas (2.25) - (2.28).

First, we optimize c and l in terms of the mean time before the result of a random blood sample is known, $\mathrm{E}[W]$ (this is, on account of Little's theorem [57], equal to $\mathrm{E}[U]/\lambda$). Therefore, we calculate $\mathrm{E}[W]$ for several values of l and c and summarize the results in table 2.1. The optimal decision variables, l_{opt} and c_{opt}, are those that correspond to the smallest value for $\mathrm{E}[W]$ in the table. This value is indicated in bold.

Analogously, we can optimize in terms of other quantities, such as the working probability (the probability that the test kit is testing during a random slot). Along the same lines, we summarize the working probability for several l's and c's in table 2.2 and indicate the best value in bold. Note that the optimal value in this case is different from the optimal value when the optimization variable was $\mathrm{E}[W]$. Minimizing a weighted sum of several parameters is also a possibility.

As the checking of all combinations of c and l can be a time-consuming assignment, we seek, in the remainder of this section, some rules of thumb that aid in determining l_{opt} and c_{opt}. Tables 2.1 and 2.2 exhibit that $\mathrm{E}[W]$ and the working probability first dramatically decrease as a function of c, then they fluctuate somewhat around the minimum, whereupon they significantly increase as c increases. Hence, one algorithm to determine (c_{opt},l_{opt}) starts with $l = c = 1$, then checks for every c each $l \leq c$, and stops when $\mathrm{E}[W]$ and the working probability drastically increase, leading to a quadratic time complexity (usually denoted by the $O(n^2)$).
We now investigate whether a small change in λ and/or p requires that the algorithm restarts from scratch. Therefore, we display (c_{opt},l_{opt}) for various values of p and λ in tables 2.3 ($\mathrm{E}[W]$ is minimized) and 2.4 (working probability is minimized). Table 2.3 shows us that, when $\mathrm{E}[W]$ has to be minimized:

- l_{opt} is small in general; one can take advantage of this observation by first searching for the best value of c corresponding to $l = 1$ ($\tilde{c}_{\text{opt}}$), and then seeking for the optimal value of l corresponding to $\tilde{c}_{\text{opt}}$. Table 2.1 shows that the resulting $\mathrm{E}[W]$ is near (or equal) to the $\mathrm{E}[W]$ corresponding to ($c_{\text{opt}}, l_{\text{opt}}$). This thus leads to an algorithm of time complexity $O(n)$ that produces near optimal values of l and c.
- l_{opt} and c_{opt} slowly increase as a function of λ. Hence, if λ increases, start at the previous optimal c-value, instead of at $c = 1$.
- l_{opt} is not influenced by p. Hence, if p alters, only search a new optimal value of c, leading to a $O(n)$ time complexity.
- If p decreases, c_{opt} increases. Hence, start searching at the previous optimal value of c.

Table 2.4 shows that, when the working probability has to be minimized:

- l_{opt}=c_{opt}. This reduces the time complexity of the algorithm from $O(n^2)$ to $O(n)$.
- λ has little to no influence on c_{opt}. Hence, if λ changes, the optimal values of l and c do not alter.
- Analogously as for $\mathrm{E}[W]$, c_{opt} increases if p decreases.

These rules of thumb thus aid in speeding up the search for a new (c_{opt},l_{opt}) when λ and/or p alters.

Remark 7. *The above rules of thumb became clear from an example with a Poisson distribution for the number of sample arrivals during a time slot. We have checked this for other distributions as well and all the above rules remain valid.*

Remark 8. *In the example, β was equal to 0. However, one can conceive β also as a decision parameter and one can thus determine a good value of β analogously as in the above examples. The purpose of setting $\beta \neq 0$ is to avoid that samples perish because their testing is postponed too long until more samples have arrived.*

Table 2.1: $\mathrm{E}[W]$ for several values of l and c whereby $\beta = 0$, $p = 0.025$ and $\lambda = 0.25$; the optimum is indicated in bold

	$c = 4$	$c = 5$	$c = 6$	$c = 7$	$\mathbf{c = 8}$	$c = 9$	$c = 10$
$l = 1$	57.4139	42.4111	37.6298	37.6303	37.5259	39.0722	40.2516
$\mathbf{l = 2}$	57.2731	42.2940	37.5278	37.5399	**37.4464**	39.0008	40.1824
$l = 3$	60.1768	45.0087	40.1156	40.1894	40.1384	41.8741	43.2097
$l = 4$	63.3439	48.1818	43.2285	43.3755	43.3763	45.2714	46.7414
$l = 5$	——	52.5499	47.4802	47.6944	47.7292	49.7700	51.3546
$l = 6$	——	——	51.1644	51.6063	51.6833	53.8279	55.4936
$l = 7$	——	——	——	56.5320	56.7008	58.9817	60.7385
$l = 8$	——	——	——	——	60.9422	63.4986	65.3284
$l = 9$	——	——	——	——	——	68.6834	70.6856
$l = 10$	——	——	——	——	——	——	75.3074

Table 2.2: Working probability for several values of l and c whereby $\beta = 0$, $p = 0.025$ and $\lambda = 0.25$; the optimum is indicated in bold

	$c = 6$	$c = 7$	$c = 8$	$c = 9$	$\mathbf{c = 10}$	$c = 11$	$c = 12$
$l = 1$	0.9796	0.9786	0.9780	0.9778	0.9777	0.9777	0.9778
$l = 2$	0.9209	0.9173	0.9150	0.9143	0.9138	0.9141	0.9143
$l = 3$	0.8475	0.8406	0.8361	0.8348	0.8338	0.8344	0.8349
$l = 4$	0.7852	0.7754	0.7689	0.7672	0.7658	0.7669	0.7677
$l = 5$	0.7436	0.7316	0.7236	0.7216	0.7200	0.72159	0.7228
$l = 6$	0.7170	0.7020	0.6928	0.6907	0.6889	0.6909	0.6926
$l = 7$	——	0.6891	0.6778	0.6756	0.6738	0.6761	0.6780
$l = 8$	——	——	0.6679	0.6653	0.6635	0.6661	0.6682
$l = 9$	——	——	——	0.6631	0.6609	0.6638	0.6660
$\mathbf{l = 10}$	——	——	——	——	**0.6591**	0.6626	0.6648
$l = 11$	——	——	——	——	——	0.6651	0.6678
$l = 12$	——	——	——	——	——	——	0.6698

Table 2.3: (c_{opt},l_{opt}) so that $\mathrm{E}\,[W]$ is minimized for several values of p and λ; * means that the test center cannot handle all the samples, because $\rho \geq 1$ for every value of c

	$p = 0.01$	$p = 0.015$	$p = 0.02$	$p = 0.025$	$p = 0.03$	$p = 0.035$
$\lambda = 0.05$	(8,1)	(7,1)	(6,1)	(6,1)	(5,1)	(4,1)
$\lambda = 0.10$	(8,1)	(7,1)	(6,1)	(6,1)	(6,1)	(4,1)
$\lambda = 0.15$	(8,1)	(7,1)	(6,1)	(6,1)	(6,1)	(6,1)
$\lambda = 0.20$	(8,1)	(8,1)	(6,1)	(6,1)	(6,1)	(6,1)
$\lambda = 0.25$	(8,2)	(8,2)	(8,2)	(8,2)	(6,2)	(6,2)
$\lambda = 0.30$	(10,2)	(8,2)	(8,2)	(8,2)	(8,2)	(8,2)
$\lambda = 0.35$	(10,2)	(8,2)	(8,2)	(8,2)	*	*

Table 2.4: (c_{opt},l_{opt}) so that the working probability is minimized for several values of p and λ; * means that the test center cannot handle all the samples, because $\rho \geq 1$ for every value of c

	$p = 0.01$	$p = 0.015$	$p = 0.02$	$p = 0.025$	$p = 0.03$	$p = 0.035$
$\lambda = 0.05$	(14,14)	(12,12)	(10,10)	(10,10)	(8,8)	(8,8)
$\lambda = 0.10$	(14,14)	(12,12)	(10,10)	(10,10)	(8,8)	(8,8)
$\lambda = 0.15$	(14,14)	(12,12)	(10,10)	(10,10)	(8,8)	(8,8)
$\lambda = 0.20$	(14,14)	(12,12)	(10,10)	(10,10)	(8,8)	(8,8)
$\lambda = 0.25$	(14,14)	(12,12)	(10,10)	(10,10)	(8,8)	(8,8)
$\lambda = 0.3$	(14,14)	(12,12)	(10,10)	(10,10)	(8,8)	(8,8)
$\lambda = 0.35$	(14,14)	(12,12)	(10,10)	(10,10)	*	*

Chapter 3

Buffer content: approximations

3.1 Preface

In this chapter, we establish, driven by numerical and interpretational motives, light- and heavy-traffic approximations for quantities (2.16)-(2.23) (sections 3.2 and 3.3). In order to calculate the buffer-related quantities (2.16)-(2.23), some numerical work is required, namely (1) the computation of the $c-1$ zeroes z_i, $i = 1, \ldots, c-1$, of $z^c - T_c(A(z))$ that are inside the closed complex unit disk and different from 1, and (2) the solution of the following set of c equations in the c unknown probabilities $d(n)$:

$$\begin{aligned}[1 - A(z_i)] \sum_{n=0}^{l-1} d(n) z_i^n + \beta \sum_{n=0}^{l-1} d(n) \left[A(z_i) z_i^n - T_n(A(z_i))\right] \\ + \sum_{n=l}^{c-1} d(n)[z_i^n - T_n(A(z_i))] = 0 \ , \qquad 1 \leq i \leq c-1 \ ,\end{aligned} \tag{3.1}$$

$$\begin{aligned}-c + \mathrm{E}\left[T_c\right] \lambda = & - c \sum_{n=0}^{l-1} d(n) + \beta \sum_{n=0}^{l-1} d(n)[c + n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]] \\ & + \sum_{n=l}^{c-1} d(n)[n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]] \ .\end{aligned} \tag{3.2}$$

The calculation of the zeroes can be a severe and even unfeasible assignment when c is large. Moreover, even when the zeroes can be computed, those zeroes might get clustered, which leads to an ill-conditioned set of equations (3.1) and (3.2) for the unknown probabilities $d(n)$. Next to the numerical motive, we are also driven by interpretational stimuli. We are eager to get insight into the behaviour of the system in the cases of light and heavy traffic.

In our paper [43], we have deduced light- and heavy-traffic approximations for the system content for the model whereby $\beta = 0$ and $T_n(z) = T_c(z), \ \forall n$. As the analysis for the more general model considered throughout this dissertation runs along the same lines as in [43], we expose the analysis for this extended model.

3.2 Light-traffic approximations

In this section, we deduce light-traffic approximations of the buffer-related quantities from chapter 2 (section 2.3). As in [8], [19], [20], [47], [62], [97], [101], we calculate light-traffic approximations by **expanding the quantities in Taylor series about $\boldsymbol{\lambda = 0}$ and retaining only the constant and linear terms** (i.e. those corresponding to λ^0 and λ^1) since the others are negligible when $\lambda \to 0$. We demonstrate this approach in detail in the next subsection, where it is applied on the system content at random slot boundaries. Thereafter, we also provide light-traffic formulas for the other buffer-related quantities, but we omit the details of the calculation, as they are analogous as for the system content at random slot marks.

3.2.1 System content at random slot boundaries

In this subsection, we establish a light-traffic approximation for the system content at random slot marks by expanding its PGF $U(\lambda, z)$ in a Taylor series about $\lambda = 0$ and retaining only the constant and the linear terms (note that we substitute every function $f(z)$ that is dependent on λ by $f(\lambda, z)$ to underline this dependency). This approach leads to a formula whereby it is required to solve a set of equations, but no zeroes have to be computed anymore.

First, recall that the PGF of the system content was established in chapter 2 (formula (2.16)) and is equal to

$$\begin{aligned} U(\lambda,z)[1-A(\lambda,z)]\left[z^c-T_c(A(\lambda,z))\right] =&(z^c-1)T_c(A(\lambda,z))[1-A(\lambda,z)]\sum_{n=0}^{l-1}d(\lambda,n)z^n \\ &+\beta\sum_{n=0}^{l-1}d(\lambda,n)g_n(\lambda,z)+\sum_{n=l}^{c-1}d(\lambda,n)h_n(\lambda,z)\ , \end{aligned} \tag{3.3}$$

with $g_n(\lambda, z)$ and $h_n(\lambda, z)$ defined as

$$\begin{aligned} g_n(\lambda,z) \triangleq&(z^n-z^c)T_n(A(\lambda,z))T_c(A(\lambda,z))+z^n(z^c-1)T_c(A(\lambda,z))A(\lambda,z) \\ &-z^c(z^n-1)T_n(A(\lambda,z))\ , \end{aligned} \tag{3.4}$$

$$\begin{aligned} h_n(\lambda,z) \triangleq&T_n(A(\lambda,z))z^c\{1-z^n-T_c(A(\lambda,z))\} \\ &-T_c(A(\lambda,z))z^n\{1-z^c-T_n(A(\lambda,z))\}\ . \end{aligned} \tag{3.5}$$

and whereby the probabilities $d(\lambda, n)$ have to be determined by solving the set of equations (3.1)-(3.2).

Next, let us denote the right-hand-side of expression (3.3) by $N_U(\lambda, z)$ (i.e. the numerator of $U(\lambda, z)$) and its Taylor series expansion about $\lambda = 0$ by $\sum_{k=0}^{\infty} N_{U,k}(z)\lambda^k$. Analogously, $D_U(\lambda, z)$ represents the denominator of $U(\lambda, z)$, i.e. $D_U(\lambda,z) = [1-A(\lambda,z)][z^c - T_c(A(\lambda,z))]$ and $\sum_{k=0}^{\infty} D_{U,k}(z)\lambda^k$ characterises its series expansion. The Taylor series of $U(\lambda,z)$ about $\lambda = 0$ can then be written as

$$\begin{aligned} U(\lambda,z) =& \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \left\{ \frac{\partial^n}{\partial\lambda^n} \left. \frac{N_U(\lambda,z)}{D_U(\lambda,z)} \right|_{\lambda=0} \right\} \\ =& \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \left\{ \frac{\partial^n}{\partial\lambda^n} \left. \frac{\sum_{k=0}^{\infty} N_{U,k}(z)\lambda^k}{\sum_{k=0}^{\infty} D_{U,k}(z)\lambda^k} \right|_{\lambda=0} \right\} \ . \end{aligned} \tag{3.6}$$

Hence, $N_{U,k}(z)$ and $D_{U,k}(z)$ have to be calculated, which, in turn, rely on the series expansions of $A(\lambda,z)$ and $T_n(A(\lambda,z))$. Let us therefore define $A^{(\mathbf{1}_n,\mathbf{2}_m)}(x,y)$, with $\mathbf{k}_n$ the series consisting of n consecutive k's, as

$$A^{(\mathbf{1}_n,\mathbf{2}_m)}(x,y) \triangleq \frac{\partial^n}{\partial\lambda^n}\frac{\partial^m}{\partial^m z} A(\lambda,z)\Big|_{\lambda=x,z=y} \ .$$

In view of this, the Taylor series expansions of $A(\lambda,z)$ and $T_n(A(\lambda,z))$ can be written as

$$\begin{aligned} A(\lambda,z) =& \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \left\{ \frac{\partial^n}{\partial\lambda^n} A(\lambda,z)\Big|_{\lambda=0} \right\} \\ =& 1 + \lambda A^{(1)}(0,z) + \frac{\lambda^2}{2} A^{(1,1)}(0,z) + O(\lambda^3) \ , \end{aligned} \tag{3.7}$$

$$\begin{aligned} T_n(A(\lambda,z)) =& 1 + \lambda \mathrm{E}\left[T_n\right] A^{(1)}(0,z) \\ &+ \frac{\lambda^2}{2}\left[T_n^{''}(1)A^{(1)}(0,z)^2 + \mathrm{E}\left[T_n\right] A^{(1,1)}(0,z)\right] + O(\lambda^3) \ . \end{aligned} \tag{3.8}$$

We have thereby taken into account that when $\lambda = 0$, no customers will ever arrive, so that $A(0,z) = 1$. Note that we have also provided the quadratic terms in λ, as these will be used later on.

Next, as (3.7) and (3.8) thus exhibit that the constant terms of $A(\lambda,z)$ and $T_n(A(\lambda,z))$ are equal to 1, we find, by examining expressions (3.3)-(3.5), that

$$N_{U,0}(z) = D_{U,0}(z) = 0 \ .$$

In addition, $N_{U,1}(z) \neq 0$ and $D_{U,1}(z) \neq 0$, so that (3.6) transforms into

$$\begin{aligned} U(\lambda,z) =& \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \left\{ \frac{\partial^n}{\partial\lambda^n} \left. \frac{N_{U,1}(z) + \lambda N_{U,2}(z) + \ldots}{D_{U,1}(z) + \lambda D_{U,2}(z) + \ldots} \right|_{\lambda=0} \right\} \\ =& \frac{N_{U,1}(z)}{D_{U,1}(z)} + \lambda \frac{N_{U,2}(z)D_{U,1}(z) - N_{U,1}(z)D_{U,2}(z)}{D_{U,1}(z)^2} + O(\lambda^2) \ . \end{aligned} \tag{3.9}$$

Further, let us denote the series expansions of $d(\lambda,n)$ (the unknowns $d(\lambda,n)$ depend on λ, due to its appearance in the set of equations (3.1)-(3.2)), $g_n(\lambda,z)$ and $h_n(\lambda,z)$ by respectively

$$d(\lambda,n) = \sum_{k=0}^{\infty} d_k(n)\lambda^k \ ,$$

$$g_n(\lambda,z) = \sum_{k=0}^{\infty} g_{n,k}(z)\lambda^k \ ,$$

$$h_n(\lambda, z) = \sum_{k=0}^{\infty} h_{n,k}(z)\lambda^k \ .$$

We then obtain that $N_{U,1}(z)$, $N_{U,2}(z)$, $D_{U,1}(z)$ and $D_{U,2}(z)$ can be written as

$$\begin{aligned}
N_{U,1}(z) =& - A^{(1)}(0,z)(z^c-1)\sum_{n=0}^{l-1} d_0(n)z^n \\
&+ \beta\sum_{n=0}^{l-1} d_0(n)g_{n,1}(z) + \sum_{n=l}^{c-1} d_0(n)h_{n,1}(z) \ , \qquad (3.10)\\
N_{U,2}(z) =& - (z^c-1)\left\{\mathrm{E}\left[T_c\right]A^{(1)}(0,z)^2 + \frac{1}{2}A^{(1,1)}(0,z)\right\}\sum_{n=0}^{l-1} d_0(n)z^n \\
&- (z^c-1)A^{(1)}(0,z)\sum_{n=0}^{l-1} d_1(n)z^n + \beta\sum_{n=0}^{l-1}\left[d_1(n)g_{n,1}(z) + d_0(n)g_{n,2}(z)\right] \\
&+ \sum_{n=l}^{c-1}\left[d_1(n)h_{n,1}(z) + d_0(n)h_{n,2}(z)\right] \ , \qquad (3.11)\\
D_{U,1}(z) =& - A^{(1)}(0,z)(z^c-1) \ , \qquad (3.12)\\
D_{U,2}(z) =& - \frac{1}{2}A^{(1,1)}(0,z)(z^c-1) + A^{(1)}(0,z)^2\mathrm{E}\left[T_c\right] \ . \qquad (3.13)
\end{aligned}$$

In addition, $g_{n,1}(z)$, $g_{n,2}(z)$, $h_{n,1}(z)$ and $h_{n,2}(z)$ are equal to

$$\begin{aligned}
g_{n,1}(z) =& A^{(1)}(0,z)\{(z^c-1)z^n(1-\mathrm{E}\left[T_n\right]) + (z^n-1)z^c\mathrm{E}\left[T_c\right]\} \ , \qquad (3.14)\\
g_{n,2}(z) =& (z^n-z^c)\mathrm{E}\left[T_n\right]\mathrm{E}\left[T_c\right]A^{(1)}(0,z)^2 \\
&+ z^n(z^c-1)\left\{\mathrm{E}\left[T_c\right]A^{(1)}(0,z)^2 + \frac{1}{2}A^{(1,1)}(0,z)\right\} \\
&- z^n(z^c-1)\frac{1}{2}\left\{T_n^{''}(1)A^{(1)}(0,z)^2 + \mathrm{E}\left[T_n\right]A^{(1,1)}(0,z)\right\} \\
&+ z^c(z^n-1)\frac{1}{2}\left\{T_c^{''}(1)A^{(1)}(0,z)^2 + \mathrm{E}\left[T_c\right]A^{(1,1)}(0,z)\right\} \ , \qquad (3.15)\\
h_{n,1}(z) =& A^{(1)}(0,z)\{z^c(z^n-1)\mathrm{E}\left[T_c\right] - z^n(z^c-1)\mathrm{E}\left[T_n\right]\} \ , \qquad (3.16)\\
h_{n,2}(z) =& \frac{1}{2}z^c(z^n-1)\left\{T_c^{''}(1)A^{(1)}(0,z)^2 + \mathrm{E}\left[T_c\right]A^{(1,1)}(0,z)\right\} \\
&- \frac{1}{2}z^n(z^c-1)\left\{T_n^{''}(1)A^{(1)}(0,z)^2 + \mathrm{E}\left[T_n\right]A^{(1,1)}(0,z)\right\} \\
&+ (z^n-z^c)\mathrm{E}\left[T_n\right]\mathrm{E}\left[T_c\right]A^{(1)}(0,z)^2 \ . \qquad (3.17)
\end{aligned}$$

Hence, in order to characterize fully the constant and linear terms of $U(\lambda, z)$, $d_0(n)$ and $d_1(n)$, the constant and linear terms of the unknown probabilities $d(n)$, have to be calculated. Remember that the probabilities $d(n)$ can be found by solving the set of equations expressing that $U(\lambda, 1) = 1$ (the normalisation condition) and that the numerator of $U(\lambda, z)$ must vanish for the $c-1$ zeroes $z_i(\lambda)$, $i = 1, \ldots, c-1$, of $z^c - T_c(A(\lambda, z))$ that are inside the closed complex unit disk and different from 1 (equations (3.1) and (3.2)). These zeroes are, in general, to be calculated numerically.

Consequently, we expand equations (3.1) and (3.2) in a Taylor series about $\lambda = 0$. Let us therefore designate the Taylor series expansion of the zeroes $z_i(\lambda)$ by $\sum_{k=0}^{\infty} z_{i,k}\lambda^k$. Expansion of equations (3.1) and (3.2) and thereby applying Newton's binomium (so that $z_i^n = z_{i,0}^n + \lambda n z_{i,1} z_{i,0}^{n-1} + O(\lambda^2)$) and taking into account that $A^{(2)}(0,z) = 0$ (since $A(0,z) = 1$) and $z_i(0) = z_{i,0}$, results in

$$\begin{cases} \beta \sum_{n=0}^{l-1} d_0(n) \left(z_{i,0}^n - 1 \right) + \sum_{n=l}^{c-1} d_0(n)(z_{i,0}^n - 1) \\ +\lambda \Big[- A^{(1)}(0, z_{i,0}) \sum_{n=0}^{l-1} d_0(n) z_{i,0}^n + \beta \sum_{n=0}^{l-1} d_1(n) \left(z_{i,0}^n - 1 \right) \\ +\beta \sum_{n=0}^{l-1} d_0(n) \left\{ A^{(1)}(0, z_{i,0}) z_{i,0}^n + n z_{i,1} z_{i,0}^{n-1} - \mathrm{E}\left[T_n\right] A^{(1)}(0, z_{i,0}) \right\} \\ + \sum_{n=l}^{c-1} d_0(n) \left\{ n z_{i,1} z_{i,0}^{n-1} - \mathrm{E}\left[T_n\right] A^{(1)}(0, z_{i,0}) \right\} \\ + \sum_{n=l}^{c-1} d_1(n) \left(z_{i,0}^n - 1 \right) \Big] + O(\lambda^2) = 0 \ , \qquad 1 \leq i \leq c-1 \ , \\ \\ -c \sum_{n=0}^{l-1} \{ d_0(n) + \lambda d_1(n) \} + \beta \sum_{n=0}^{l-1} \{ d_0(n) + \lambda d_1(n) \} \{ c + n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right] \} \\ + \sum_{n=l}^{c-1} \{ d_0(n) + \lambda d_1(n) \} \{ n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right] \} = -c + \mathrm{E}\left[T_c\right] \lambda \ . \end{cases} \tag{3.18}$$

Hence, in order to calculate $d_0(n)$ and $d_1(n)$ from (3.18), we still have to deduce the constant $(z_{i,0})$ and linear $(z_{i,1})$ terms of the Taylor series expansion of the zeroes $z_i(\lambda)$, $i = 1, \ldots, c-1$. To this end, we expand both sides of $z_i(\lambda)^c = T_c(A(\lambda, z_i(\lambda)))$ in a Taylor series about $\lambda = 0$. We thereby apply Newton's binomium and we take into account that $z_i(0) = z_{i,0}$ and $A^{(2)}(0, z_{i,0}) = 0$. As a result, $z_i(\lambda)^c = T_c(A(\lambda, z_i(\lambda)))$ is transformed into

$$z_{i,0}^c + \lambda c z_{i,0}^{c-1} z_{i,1} + O(\lambda^2) = 1 + \lambda \mathrm{E}\left[T_c\right] A^{(1)}\left(0, z_{i,0}\right) + O(\lambda^2) \ .$$

Equating the constant term at the left-hand-side with the constant term at the right-hand-side and repeating this for the linear terms produces:

$$\begin{cases} z_{i,0}^c = 1 \ , \\ c z_{i,0}^{c-1} z_{i,1} = \mathrm{E}\left[T_c\right] A^{(1)}\left(0, z_{i,0}\right) \ . \end{cases}$$

It is directly clear that the first equation has c solutions, the c complex c-th roots of one: $\varepsilon_i \triangleq e^{(\imath 2\pi i)/c}$, with $\imath$ the imaginary unit and $i = 0, \ldots, c-1$[1]. Hence,

$$z_{i,0} = \varepsilon_i \ , \qquad 0 \leq i \leq c-1 \ . \tag{3.19}$$

The corresponding $z_{i,1}$'s can be found by replacing $z_{i,0}$ by ε_i (and thus $z_{i,0}^c$ by 1) in the second equation, leading to

$$z_{i,1} = \frac{\varepsilon_i \mathrm{E}\left[T_c\right]}{c} A^{(1)}\left(0, \varepsilon_i\right) \ , \qquad 0 \leq i \leq c-1 \ . \tag{3.20}$$

This concludes the calculation of $z_{i,0}$ and $z_{i,1}$ $(i = 0, \ldots, c-1)$. Note that these are exactly known.

We now rely on this result in the calculation of $d_0(n)$ and $d_1(n)$. Substituting

[1] Note that $z_{0,0} = \varepsilon_0 = 1$, which is logical because one of the c zeroes z_i of $z^c - T_c(A(z))$ is equal to one. This zero was in chapter 2 denoted by z_0.

(3.19) and (3.20) in (3.18) yields

$$\begin{cases} \beta \sum_{n=1}^{l-1} d_0(n)\left(\varepsilon_i^n - 1\right) + \sum_{n=l}^{c-1} d_0(n)(\varepsilon_i^n - 1) \\ +\lambda\Big[- A^{(1)}(0,\varepsilon_i) \sum_{n=0}^{l-1} d_0(n)\varepsilon_i^n + \beta \sum_{n=1}^{l-1} d_1(n)\left(\varepsilon_i^n - 1\right) \\ +\beta A^{(1)}(0,\varepsilon_i) \sum_{n=0}^{l-1} d_0(n)\left\{\varepsilon_i^n + n\varepsilon_i^n \mathrm{E}\left[T_c\right]/c - \mathrm{E}\left[T_n\right]\right\} \\ +A^{(1)}(0,\varepsilon_i) \sum_{n=l}^{c-1} d_0(n)\left\{n\varepsilon_i^n \mathrm{E}\left[T_c\right]/c - \mathrm{E}\left[T_n\right]\right\} \\ + \sum_{n=l}^{c-1} d_1(n)\left(\varepsilon_i^n - 1\right)\Big] + O(\lambda^2) = 0 \ , \qquad 1 \le i \le c-1 \ , \\ \\ -c \sum_{n=0}^{l-1} \{d_0(n) + \lambda d_1(n)\} + \beta \sum_{n=0}^{l-1} \{d_0(n) + \lambda d_1(n)\}\{c + n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]\} \\ + \sum_{n=l}^{c-1} \{d_0(n) + \lambda d_1(n)\}\{n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]\} = -c + \mathrm{E}\left[T_c\right]\lambda \ . \end{cases} \tag{3.21}$$

Equating the constant terms at the left-hand-sides with the constant terms at the right-hand-sides yields:

$$\beta \sum_{n=1}^{l-1} d_0(n)\left(\varepsilon_i^n - 1\right) + \sum_{n=l}^{c-1} d_0(n)(\varepsilon_i^n - 1) = 0 \ , \qquad 1 \le i \le c-1 \ , \tag{3.22a}$$

$$-c \sum_{n=0}^{l-1} d_0(n) + \beta \sum_{n=0}^{l-1} d_0(n)\{c + n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]\} + \sum_{n=l}^{c-1} d_0(n)\{n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]\} = -c \ . \tag{3.22b}$$

We now make a distinction between $\beta \neq 0$ and $\beta = 0$.

case 1: $\beta \neq 0$

In this case, (3.22a) forms a set of $c-1$ independent linear equations in $c-1$ unknowns $d_0(n)$, $n = 1, \ldots c-1$. It is clear that

$$d_0(n) = 0 \ , \qquad 1 \le n \le c-1 \ ,$$

is a solution of the set of equations. Furthermore, one can prove that this solution is unique. As a result, (3.22b) produces:

$$d_0(0) = \frac{1}{1 - \beta(1 - \mathrm{E}\left[T_0\right])} \ . \tag{3.23}$$

This can be explained intuitively. When the number of arriving customers goes to zero ($\lambda \to 0$), no customers will ever be in the system, because the few customers that arrive are served after a geometrically distributed time (each slot a service starts with probability β), and this time is negligible as compared to the nearly infinite interarrival times. Therefore, - recall definition (2.4) for $d(n)$ ($d(n) = \lim_{k\to\infty} \Pr\left[Q_k + A_k = n, R_k \le 1\right]$) - $d_0(0) = \Pr\left[R \le 1\right]$. On the other hand, the system being virtually always empty implies that the system alternates between periods whereby the server processes 0 customers (with mean length $\mathrm{E}\left[T_0\right]$) and periods whereby the server is not processing (with mean length $(1-\beta)/\beta$). Hence, the fraction of slots during which the server is not processing (i.e. $R = 0$) or during which the server is in the last slot of service ($R = 1$) equals

$$\frac{(1-\beta)/\beta}{\mathrm{E}\left[T_0\right] + (1-\beta)/\beta} + \frac{1}{\mathrm{E}\left[T_0\right] + (1-\beta)/\beta} \ ,$$

which is equal to expression (3.23).

Next, we equate the linear terms of (3.21) and on account of $d_0(n) = 0$, $n =$

$1, \ldots, c-1$ and (3.23), we obtain the following set of equations in $d_1(n)$, $n = 0, \ldots, c-1$:

$$\beta \sum_{n=1}^{l-1} d_1(n)\,(\varepsilon_i^n - 1) + \sum_{n=l}^{c-1} d_1(n)\,(\varepsilon_i^n - 1) = A^{(1)}(0, \varepsilon_i) \ , \qquad 1 \le i \le c-1 \ ,$$

$$\begin{aligned} &- c \sum_{n=0}^{l-1} d_1(n) + \beta \sum_{n=0}^{l-1} d_1(n)\{c + n\mathrm{E}\,[T_c] - c\mathrm{E}\,[T_n]\} \\ &+ \sum_{n=l}^{c-1} d_1(n)\{n\mathrm{E}\,[T_c] - c\mathrm{E}\,[T_n]\} = \mathrm{E}\,[T_c] \ . \end{aligned} \tag{3.24}$$

Hence, (3.24) provides c equations from which the c unknowns $d_1(n)$ ($n = 0, \ldots, c-1$) can be calculated. It can again be proved that (3.24) has a unique solution. Finally, the combination of (3.9)-(3.17), $d_0(n) = 0$ for $n = 1, \ldots, c-1$, and expression (3.23) for $d_0(0)$ yields the light-traffic approximation for $U(\lambda, z)$ in the case $\beta \neq 0$:

$$\begin{aligned} U(\lambda, z) = 1 + \lambda \frac{1}{z^c - 1} \cdot \Bigg[& z^c \mathrm{E}\,[T_c]\, A^{(1)}(0, z) + \frac{1}{2}\beta \frac{(z^c - 1) T_0''(1) A^{(1)}(0, z)}{1 - \beta(1 - \mathrm{E}\,[T_0])} \\ & + (z^c - 1) \sum_{n=0}^{l-1} d_1(n) z^n - \beta(z^c - 1) \sum_{n=0}^{l-1} d_1(n) z^n \\ & + \beta \sum_{n=0}^{l-1} d_1(n) f_n(z) + \sum_{n=l}^{c-1} d_1(n) f_n(z) \Bigg] \\ & + O(\lambda^2) \ , \end{aligned} \tag{3.25}$$

with

$$f_n(z) \triangleq z^n (z^c - 1)\mathrm{E}\,[T_n] - z^c (z^n - 1)\mathrm{E}\,[T_c] \ .$$

Formula (3.25) thus reveals a.o. that

$$\lim_{\lambda \to 0} \Pr\,[U = n] = \begin{cases} 1 & \text{if } n = 0 \ , \\ 0 & \text{else} \ . \end{cases}$$

This can be explained as follows: when customers arrive for small λ, they will be served after some relatively short time (because $\beta \neq 0$) and this period is negligible as compared to the nearly infinite time until new customers arrive. As a result, the system is virtually always empty (note that this confirms the results from Fig. 2.2(a)).

Remark 9. *We have implicitly assumed that $\beta >> \lambda$ in this subsection. Otherwise, the first sum in (3.22a) would be negligible as compared to the second, so that it cannot be concluded anymore that $d_0(n) = 0$, $1 \le n \le l-1$. When $\beta >> \lambda$ does not hold, we advise to use the approximation formulas corresponding to the case $\beta = 0$.*

Remark 10. *Although it is implicitly assumed that $l \ge 1$ in (3.23), results for $l = 0$ can be obtained from light-traffic formulas (3.23) and (3.25) by setting $\beta = 1$. Indeed, a system with $\beta = 1$ is, regardless of l, equivalent with a system whereby $l = 0$. In both cases, the server always initiates a new service, even when no customers are available.*

case 2: $\beta = 0$

Equation (3.22a) transforms into:

$$\sum_{n=l}^{c-1} d_0(n)(\varepsilon_i^n - 1) = 0 \ , \qquad 1 \leq i \leq c-1 \ . \tag{3.26}$$

As the first sum of (3.22a) has vanished, this equation provides no information about $d_0(n)$ for $n < l$ anymore. We still can conclude from (3.26), however, that

$$d_0(n) = 0 \ , \qquad l \leq n \leq c-1 \ . \tag{3.27}$$

As a result, and because $\beta = 0$, equation (3.22b) becomes

$$\sum_{n=0}^{l-1} d_0(n) = 1 \ . \tag{3.28}$$

Equating the linear terms from (3.21) and accounting for (3.27) leads to the following equations:

$$-A^{(1)}(0,\varepsilon_i)\sum_{n=0}^{l-1} d_0(n)\varepsilon_i^n + \sum_{n=l}^{c-1} d_1(n)\left(\varepsilon_i^n - 1\right) = 0 \ , \qquad 1 \leq i \leq c-1 \ , \tag{3.29a}$$

$$-c\sum_{n=0}^{l-1} d_1(n) + \sum_{n=l}^{c-1} d_1(n)\left\{n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]\right\} = \mathrm{E}\left[T_c\right] \ . \tag{3.29b}$$

Expressions (3.28), (3.29a) and (3.29b) produce $c+1$ equations in $c+l$ unknowns $d_0(n)$, $n = 0, \ldots, l-1$, and $d_1(n)$, $n = 0, \ldots, c-1$. Hence, when l is larger than 1, extra equations are required. These are provided by equalising the quadratic terms of (3.1) and taking into account expressions (3.19) and (3.20) for $z_{i,0}$ and $z_{i,1}$:

$$\begin{aligned}
&-\frac{1}{2}\left\{A^{(1,1)}(0,\varepsilon_i) + 2A^{(1,2)}(0,\varepsilon_i)\frac{\varepsilon_i}{c}\mathrm{E}\left[T_c\right]A^{(1)}(0,\varepsilon_i)\right\}\sum_{n=0}^{l-1} d_0(n)\varepsilon_i^n \\
&- A^{(1)}(0,\varepsilon_i)\left\{\sum_{n=0}^{l-1} d_1(n)\varepsilon_i^n + \sum_{n=0}^{l-1} d_0(n)\varepsilon_i^n n\frac{\mathrm{E}\left[T_c\right]}{c}A^{(1)}(0,\varepsilon_i)\right\} \\
&+ \sum_{n=l}^{c-1} d_2(n)(\varepsilon_i^n - 1) + A^{(1)}(0,\varepsilon_i)\sum_{n=l}^{c-1} d_1(n)\left\{n\varepsilon_i^n\mathrm{E}\left[T_c\right]/c - \mathrm{E}\left[T_n\right]\right\} = 0 \ , \qquad 1 \leq i \leq c-1 \ .
\end{aligned} \tag{3.30}$$

The unknowns $d_0(n)$ and $d_1(n)$ can now be found by solving equations (3.28)-(3.30), which together produce a set of $2c$ equations in $2c$ unknowns. As a bonus, we have also found $d_2(n)$ $(l \leq n \leq c-1)$. The combination of (3.9)-(3.17) and $d_0(n) = 0$, $n = l, \ldots, c-1$, yields the light-traffic approximation for $U(\lambda, z)$ in the case $\beta = 0$:

$$\begin{aligned}
U(\lambda, z) = \sum_{n=0}^{l-1} d_0(n)z^n + \lambda\frac{1}{z^c - 1}\cdot\Bigg[&z^c\mathrm{E}\left[T_c\right]A^{(1)}(0,z)\sum_{n=0}^{l-1} d_0(n)z^n \\
&+ (z^c - 1)\sum_{n=0}^{l-1} d_1(n)z^n + \sum_{n=l}^{c-1} d_1(n)f_n(z)\Bigg] \\
&+ O(\lambda^2) \ ,
\end{aligned} \tag{3.31}$$

whence

$$\lim_{\lambda\to 0}\Pr\left[U=n\right]=\begin{cases} d_0(n) & \text{if } 0\le n\le l-1 \ , \\ 0 & \text{else} \ . \end{cases}$$

This result can also be explained intuitively. If customers arrive and after arrival the system content is smaller than l, the customers remain in the queue. This implies that the buffer is non-zero for a non-negligible time, thus that $\Pr\left[U>0\right]\neq 0$ when $\beta=0$ and $l>1$ (this can also be observed in Fig. 2.2(a)). When the system content reaches (or exceeds) l, up to c customers are served. Since service times are negligible as compared to interarrival times, the time that l or more customers are in the system tends to zero.

Remark 11. *As we have implicitly assumed that $l\ge 1$ in (3.27), and since $\beta=0$, light-traffic approximation (3.31) is not valid for $l=0$. As mentioned in remark 10, the light-traffic formula for $l=0$ can be obtained by setting $\beta=1$ in (3.25). This remark is also valid when considering light-traffic approximations for other quantities below.*

3.2.2 Queue content at random slot boundaries

We now show light-traffic approximations for other random variables calculated in chapter 2. Let us start with the PGF $Q(\lambda,z)$ of the queue content at random slot boundaries.

We denote the Taylor series expansion of the numerator and the denominator of $Q(\lambda,z)$ by respectively $\sum_{k=0}^{\infty}N_{Q,k}(z)\lambda^k$ and $\sum_{k=0}^{\infty}D_{Q,k}(z)\lambda^k$. Analogously as in subsection 3.2.1, we have that $N_{Q,0}(z)=D_{Q,0}(z)=0$, so that

$$Q(\lambda,z)=\frac{N_{Q,1}(z)}{D_{Q,1}(z)}+\lambda\frac{N_{Q,2}(z)D_{Q,1}(z)-N_{Q,1}(z)D_{Q,2}(z)}{D_{Q,1}(z)^2}+O(\lambda^2) \ .$$

Along the same lines as in the previous subsection for $U(\lambda,z)$, we obtain the following light-traffic approximation for $Q(\lambda,z)$ when $\beta\neq 0$:

$$\begin{aligned} Q(\lambda,z)=1+\lambda\frac{1}{z^c-1}\cdot\Bigg[&A^{(1)}(0,z)\mathrm{E}\left[T_c\right]+\frac{1}{2}\beta\frac{(z^c-1)T_0^{''}(1)A^{(1)}(0,z)}{1-\beta(1-\mathrm{E}\left[T_0\right])} \\ &+(z^c-1)\sum_{n=0}^{l-1}d_1(n)z^n-\beta(z^c-1)\sum_{n=0}^{l-1}d_1(n)z^n \\ &+\beta\sum_{n=0}^{l-1}d_1(n)\tilde{f}_n(z)+\sum_{n=l}^{c-1}d_1(n)\tilde{f}_n(z)\Bigg] \\ &+O(\lambda^2) \ , \end{aligned} \tag{3.32}$$

with

$$\tilde{f}_n(z)\triangleq(z^c-1)\mathrm{E}\left[T_n\right]-(z^n-1)\mathrm{E}\left[T_c\right] \ ,$$

whereas

$$\begin{aligned} Q(\lambda,z)=\sum_{n=0}^{l-1}d_0(n)z^n+\lambda\frac{1}{z^c-1}\cdot\Bigg[&A^{(1)}(0,z)\mathrm{E}\left[T_c\right]\sum_{n=0}^{l-1}d_0(n)z^n \\ &+(z^c-1)\sum_{n=0}^{l-1}d_1(n)z^n+\sum_{n=l}^{c-1}d_1(n)\tilde{f}_n(z)\Bigg] \\ &+O(\lambda^2) \ , \end{aligned} \tag{3.33}$$

when $\beta = 0$. Formulas (3.32) and (3.33) allow us to easily extract the constant term of the probabilities $\Pr[Q = n]$:

$$\lim_{\lambda \to 0} \Pr[Q = n] = \begin{cases} 1 & \text{if } n = 0 \ , \\ 0 & \text{else} \ , \end{cases}$$

when $\beta \neq 0$ and

$$\lim_{\lambda \to 0} \Pr[Q = n] = \begin{cases} d_0(n) & \text{if } 0 \leq n \leq l - 1 \ , \\ 0 & \text{else} \ , \end{cases}$$

if $\beta = 0$. Note that these constant terms are equal to those of $\Pr[U = n]$. Indeed, as we have explained that service times are negligible as compared to interarrival times, the server is nearly always empty. Hence, the (potential) customers in the system are all waiting in the queue.

3.2.3 System content at service completion times

In order to calculate the light-traffic approximation for the system content at service completion times, the light-traffic approximation of $F(\lambda, 1, 1)$ is required. Completely analogously as above, we find that the series expansion about $\lambda = 0$ of expression (2.19) for $F(\lambda, 1, 1)$ reads:

$$F(\lambda,1,1) = \frac{\beta}{1-\beta(1-\mathrm{E}[T_0])} + \lambda \frac{1}{c} \cdot \left[1 + \beta \sum_{n=0}^{l-1} d_1(n)(c-n) \right. \\ \left. + \sum_{n=l}^{c-1} d_1(n)(c-n) \right] + O(\lambda^2) \ ,$$

when $\beta \neq 0$, whereas in the case $\beta = 0$, we obtain

$$F(\lambda,1,1) = \lambda \frac{1 + \sum_{n=l}^{c-1} d_1(n)(c-n)}{c} + \lambda^2 \frac{\sum_{n=l}^{c-1} d_2(n)(c-n)}{c} + O(\lambda^3) \ .$$

Note that we have also computed the quadratic term of $F(\lambda, 1, 1)$ in the latter case. The reason is that we need this term in the calculation of the PGF $\tilde{U}(\lambda, z)$ of the system content at service completion times.

Let us now continue with $\tilde{U}(\lambda, z)$. Therefore, we denote the series expansions of its numerator and denominator respectively by $\sum_{k=0}^{\infty} N_{\tilde{U},k}(z)\lambda^k$ and $\sum_{k=0}^{\infty} D_{\tilde{U},k}(z)\lambda^k$ and we designate the constant, linear and quadratic terms of $F(\lambda, 1, 1)$ by respectively $F_0(1, 1)$, $F_1(1, 1)$ and $F_2(1, 1)$. When $\beta \neq 0$, we obtain that $N_{\tilde{U},0}(z) = D_{\tilde{U},0}(z) \neq 0$, and consequently, that

$$\tilde{U}(\lambda, z) = 1 + \lambda \frac{N_{\tilde{U},1}(z) - D_{\tilde{U},1}(z)}{D_{\tilde{U},0}(z)} + O(\lambda^2) \ ,$$

with

$$\begin{aligned}
N_{\tilde{U},1}(z) =& \frac{A^{(1)}(0,z)}{1-\beta(1-\mathrm{E}\left[T_0\right])} + \beta \frac{A^{(1)}(0,z)}{1-\beta(1-\mathrm{E}\left[T_0\right])} \left\{z^c \mathrm{E}\left[T_0\right] - \mathrm{E}\left[T_c\right] - 1\right\} \\
&+ \beta \sum_{n=0}^{l-1} d_1(n)(z^c - z^n) + \sum_{n=l}^{c-1} d_1(n)(z^c - z^n) \ , \\
D_{\tilde{U},0}(z) =& F_0(1,1)(z^c-1) = \beta \frac{(z^c-1)}{1-\beta(1-\mathrm{E}\left[T_0\right])} \ , \\
D_{\tilde{U},1}(z) =& F_1(1,1)(z^c-1) - F_0(1,1)\mathrm{E}\left[T_c\right] A^{(1)}(0,z) \\
=& \frac{(z^c-1)}{c} \left[1 + \beta \sum_{n=0}^{l-1} d_1(n)(c-n) + \sum_{n=l}^{c-1} d_1(n)(c-n)\right] \\
&- \mathrm{E}\left[T_c\right] A^{(1)}(0,z) \frac{\beta}{1-\beta(1-\mathrm{E}\left[T_0\right])} \ .
\end{aligned}$$

If, on the other hand, $\beta = 0$, it holds that $N_{\tilde{U},0}(z) = D_{\tilde{U},0}(z) = 0$, and consequently, we obtain

$$\tilde{U}(\lambda, z) = \frac{N_{\tilde{U},1}(z)}{D_{\tilde{U},1}(z)} + \lambda \frac{N_{\tilde{U},2}(z) D_{\tilde{U},1}(z) - N_{\tilde{U},1}(z) D_{\tilde{U},2}(z)}{D_{\tilde{U},1}(z)^2} + O(\lambda^2) \ ,$$

with

$$\begin{aligned}
N_{\tilde{U},1}(z) =& A^{(1)}(0,z) \sum_{n=0}^{l-1} d_0(n) z^n + \sum_{n=l}^{c-1} d_1(n)(z^c - z^n) \ , \\
N_{\tilde{U},2}(z) =& \mathrm{E}\left[T_c\right] A^{(1)}(0,z)^2 \sum_{n=0}^{l-1} d_0(n) z^n + \frac{1}{2} A^{(1,1)}(0,z) \sum_{n=0}^{l-1} d_0(n) z^n \\
&+ A^{(1)}(0,z) \sum_{n=0}^{l-1} d_1(n) z^n + \sum_{n=l}^{c-1} d_2(n)(z^c - z^n) \\
&+ \sum_{n=l}^{c-1} d_1(n) A^{(1)}(0,z)(z^c \mathrm{E}\left[T_n\right] - z^n \mathrm{E}\left[T_c\right]) \ , \\
D_{\tilde{U},1}(z) =& F_1(1,1)(z^c-1) = \frac{(z^c-1)}{c} \left[1 + \sum_{n=l}^{c-1} d_1(n)(c-n)\right] \ , \\
D_{\tilde{U},2}(z) =& F_2(1,1)(z^c-1) - F_1(1,1)\mathrm{E}\left[T_c\right] A^{(1)}(0,z) \\
=& \frac{(z^c-1)}{c} \sum_{n=l}^{c-1} d_2(n)(c-n) - \frac{\mathrm{E}\left[T_c\right] A^{(1)}(0,z)}{c} \left[1 + \sum_{n=l}^{c-1} d_1(n)(c-n)\right] \ .
\end{aligned}$$

3.2.4 Server content at random slot boundaries

We now continue with light-traffic approximation formulas for the server content S (with PGF $S(\lambda, z)$) at random slot marks. As the approach runs along the same lines as in section 3.2.1, we opt to briefly mention the results.
Let us designate the series expansions of the numerator and denominator of $S(\lambda, z)$ by respectively $\sum_{k=0}^{\infty} N_{S,k}(z)\lambda^k$ and $\sum_{k=0}^{\infty} D_{S,k}(z)\lambda^k$. The Taylor series of $S(\lambda, z)$ then reads

$$S(\lambda, z) = \frac{N_{S,0}(z)}{D_{S,0}(z)} + \lambda \frac{N_{S,1}(z) D_{S,0}(z) - N_{S,0}(z) D_{S,1}(z)}{D_{S,0}(z)^2} + O(\lambda^2) \ .$$

After some calculations, we find

$$\begin{aligned} S(\lambda,z) = 1 + \lambda\frac{1}{c}\cdot\Bigg[& z^c \mathrm{E}\left[T_c\right] + (1-\beta)c\sum_{n=0}^{l-1} d_1(n) \\ & + \beta c\sum_{n=0}^{l-1} d_1(n)z^n\mathrm{E}\left[T_n\right] + c\sum_{n=l}^{c-1} d_1(n)z^n\mathrm{E}\left[T_n\right] \\ & - z^c\mathrm{E}\left[T_c\right]\beta\sum_{n=0}^{l-1} d_1(n)n - z^c\mathrm{E}\left[T_c\right]\sum_{n=l}^{c-1} d_1(n)n\Bigg] + O(\lambda^2)\ , \end{aligned} \tag{3.34}$$

when $\beta \neq 0$, whereas in the case $\beta = 0$, we obtain

$$\begin{aligned} S(\lambda,z) = 1 + \lambda\frac{1}{c}\cdot\Bigg[& c\sum_{n=0}^{l-1} d_1(n) + c\sum_{n=l}^{c-1} d_1(n)z^n\mathrm{E}\left[T_n\right] \\ & + z^c\mathrm{E}\left[T_c\right] - z^c\mathrm{E}\left[T_c\right]\sum_{n=l}^{c-1} d_1(n)n\Bigg] + O(\lambda^2)\ . \end{aligned} \tag{3.35}$$

From formulas (3.34) and (3.35), it is straightforward to extract the constant and linear terms of the corresponding probabilities:

$$\Pr\left[S=n\right] =$$
$$\begin{cases} 1+\lambda\left[(1-\beta)\sum_{m=0}^{l-1} d_1(m) + \beta d_1(0)\mathrm{E}\left[T_0\right]\right] + O\left(\lambda^2\right) & \text{if } n=0\ , \\ \lambda\beta\mathrm{E}\left[T_n\right]d_1(n) + O\left(\lambda^2\right) & \text{if } 1\leq n\leq l-1\ , \\ \lambda\mathrm{E}\left[T_n\right]d_1(n) + O\left(\lambda^2\right) & \text{if } l\leq n\leq c-1\ , \\ \lambda\frac{\mathrm{E}[T_c]}{c}\left[1-\beta\sum_{m=0}^{l-1} d_1(m)m - \sum_{m=l}^{c-1} d_1(m)m\right] + O\left(\lambda^2\right) & \text{if } n=c\ , \\ 0 & \text{else}\ , \end{cases}$$

if $\beta \neq 0$ and

$$\Pr\left[S=n\right] = \begin{cases} 1+\lambda\sum_{m=0}^{l-1} d_1(m) + O\left(\lambda^2\right) & \text{if } n=0\ , \\ \lambda\mathrm{E}\left[T_n\right]d_1(n) + O\left(\lambda^2\right) & \text{if } l\leq n\leq c-1\ , \\ \lambda\frac{\mathrm{E}[T_c]}{c}\left[1-\sum_{m=l}^{c-1} d_1(m)m\right] + O\left(\lambda^2\right) & \text{if } n=c\ , \\ 0 & \text{else}\ , \end{cases}$$

when $\beta = 0$. The constant term being equal to zero for $n > 0$ is no surprise because service times are negligible as compared to interarrival times.

3.2.5 Number of customers in a served batch

Along the same lines as above, we establish the following light-traffic approximation for the PGF $\tilde{S}(\lambda, z)$ of the number of customers in a served batch, when $\beta \neq 0$:

$$\tilde{S}(\lambda,z) = 1 + \lambda\frac{N_{\tilde{S},1}(z) - D_{\tilde{S},1}(z)}{D_{\tilde{S},0}(z)} + O(\lambda^2)\ , \tag{3.36}$$

with

$$\begin{aligned}
N_{\tilde{S},1}(z) =& z^c - \beta\frac{\mathrm{E}\left[T_c\right]}{1-\beta(1-\mathrm{E}\left[T_0\right])} - \beta z^c\sum_{n=0}^{l-1} d_1(n)n \\
& - z^c\sum_{n=l}^{c-1} d_1(n)n + \beta c\sum_{n=0}^{l-1} d_1(n)z^n + c\sum_{n=l}^{c-1} d_1(n)z^n \ , \\
D_{\tilde{S},0}(z) =& F_0(1,1)c = \beta\frac{c}{1-\beta(1-\mathrm{E}\left[T_0\right])} \ , \\
D_{\tilde{S},1}(z) =& cF_1(1,1) - \mathrm{E}\left[T_c\right]F_0(1,1) \\
=& 1 + \beta\sum_{n=0}^{l-1} d_1(n)(c-n) + \sum_{n=l}^{c-1} d_1(n)(c-n) - \mathrm{E}\left[T_c\right]\frac{\beta}{1-\beta(1-\mathrm{E}\left[T_0\right])} \ .
\end{aligned}$$

If, on the other hand, $\beta = 0$, the approximation formula reads:

$$\tilde{S}(\lambda, z) = \frac{N_{\tilde{S},1}(z)}{D_{\tilde{S},1}(z)} + \lambda\frac{N_{\tilde{S},2}(z)D_{\tilde{S},1}(z) - N_{\tilde{S},1}(z)D_{\tilde{S},2}(z)}{D_{\tilde{S},1}(z)^2} + O(\lambda^2) \ , \tag{3.37}$$

with

$$\begin{aligned}
N_{\tilde{S},1}(z) =& z^c + \sum_{n=l}^{c-1} d_1(n)[cz^n - nz^c] \ , \\
N_{\tilde{S},2}(z) =& z^c\sum_{n=0}^{l-1} d_1(n) - z^c\sum_{n=l}^{c-1} d_2(n)n + z^c\sum_{n=l}^{c-1} d_1(n)\mathrm{E}\left[T_n\right] \\
& + c\sum_{n=l}^{c-1} d_2(n)z^n - \mathrm{E}\left[T_c\right]\sum_{n=l}^{c-1} d_1(n)z^n \ , \\
D_{\tilde{S},1}(z) =& F_1(1,1)c = 1 + \sum_{n=l}^{c-1} d_1(n)(c-n) \ , \\
D_{\tilde{S},2}(z) =& cF_2(1,1) - \mathrm{E}\left[T_c\right]F_1(1,1) \\
=& \sum_{n=l}^{c-1} d_2(n)(c-n) - \frac{\mathrm{E}\left[T_c\right]}{c}\left\{1 + \sum_{n=l}^{c-1} d_1(n)(c-n)\right\} \ .
\end{aligned}$$

As a result, the constant term of the corresponding probabilities $\Pr\left[\tilde{S} = n\right]$ reads:

$$\lim_{\lambda\to 0}\Pr\left[\tilde{S} = n\right] = \begin{cases} 1 & \text{if } n = 0 \ , \\ 0 & \text{else} \ , \end{cases}$$

if $\beta \neq 0$, whereas when $\beta = 0$

$$\lim_{\lambda\to 0}\Pr\left[\tilde{S} = n\right] = \begin{cases} \frac{d_1(n)c}{1+\sum_{m=l}^{c-1} d_1(m)(c-m)} & \text{if } l \leq n \leq c-1 \ , \\ \frac{1-\sum_{m=l}^{c-1} d_1(m)m}{1+\sum_{m=l}^{c-1} d_1(m)(c-m)} & \text{if } n = c \ , \\ 0 & \text{else} \ . \end{cases}$$

In the latter case, it is evident that the server processes batches of minimum size l and maximum size c, because no service can be initiated before at least l customers are present. When, on the other hand, $\beta \neq 0$, the server regularly starts a new service anyway. On account of the nearly infinite interarrival times, the served batches are practically always empty.

3.2.6 Probability that the server processes

The probability that the server processes a batch during a random slot is found by invoking that $q_0(\lambda, n) = d(\lambda, n)(1-\beta)$ and substituting $d(\lambda, n)$ by its Taylor series expansion, resulting in:

$$\Pr\left[\text{server processes}\right] = \frac{\beta \mathrm{E}\left[T_0\right]}{1-\beta+\beta \mathrm{E}\left[T_0\right]} - \lambda(1-\beta)\sum_{n=0}^{l-1} d_1(n) + O\left(\lambda^2\right) \quad ,$$

when $\beta \neq 0$ and

$$\Pr\left[\text{server processes}\right] = -\lambda \sum_{n=0}^{l-1} d_1(n) + O\left(\lambda^2\right) \quad ,$$

if $\beta = 0$. The constant term being zero in the latter case is again a consequence of the negligible service times. When, on the other hand $\beta \neq 0$, the system nearly always alternates between periods whereby the server processes 0 customers (with mean length $\mathrm{E}\left[T_0\right]$) and periods whereby the server is not serving (mean length $(1-\beta)/\beta$).

3.2.7 Queue content when the server not processes

Finally, we establish a light-traffic approximation for the PGF $\tilde{Q}(\lambda, z)$ of the queue content when the server not processes. We find analogously as above that the light-traffic approximation reads

$$\tilde{Q}(\lambda, z) = 1 + \lambda[1-\beta(1-\mathrm{E}\left[T_0\right])]\left[\sum_{n=1}^{l-1} d_1(n)z^n - \sum_{n=1}^{l-1} d_1(n)\right] + O\left(\lambda^2\right) \quad ,$$

when $\beta \neq 0$, whence it follows that

$$\Pr\left[\tilde{Q} = n\right] = \begin{cases} 1 - \lambda\left[1-\beta(1-\mathrm{E}\left[T_0\right])\right]\sum_{m=1}^{l-1} d_1(m) + O\left(\lambda^2\right) & \text{if } n = 0 \ , \\ \lambda\left[1-\beta(1-\mathrm{E}\left[T_0\right])\right] d_1(n) + O\left(\lambda^2\right) & \text{if } 1 \leq n \leq l-1 \ , \\ 0 & \text{else } . \end{cases}$$

On the other hand, if $\beta = 0$, we deduce that

$$\tilde{Q}(\lambda, z) = \sum_{n=0}^{l-1} d_0(n)z^n + \lambda\left[\sum_{n=0}^{l-1} d_1(n)z^n - \left\{\sum_{n=0}^{l-1} d_0(n)z^n\right\}\sum_{m=0}^{l-1} d_1(m)\right] + O\left(\lambda^2\right) \quad ,$$

which yields

$$\Pr\left[\tilde{Q} = n\right] = \begin{cases} d_0(n) + \lambda\left[d_1(n) - d_0(n)\sum_{m=0}^{l-1} d_1(m)\right] + O\left(\lambda^2\right) & \text{if } 0 \leq n \leq l-1 \ , \\ 0 & \text{else } . \end{cases}$$

The probabilities can be explained intuitively analogously as for $\Pr\left[Q = n\right]$.

Remark 12. *Note that $d(\lambda, n)$, z_i and all quantities $U(\lambda, z)$, $Q(\lambda, z)$ et cetera, can only be expanded in a Taylor series if they are analytic at $\lambda = 0$. In appendix A, we prove that if $A(\lambda, z)$ is analytic at $\lambda = 0$ for all z in the closed complex unit disk (this assumption is for instance fulfilled for the Bernoulli, geometric and Poisson distributions), then these functions are also analytic at $\lambda = 0$.*

Remark 13. *From the light-traffic approximations of the PGFs $U(\lambda, z)$, $Q(\lambda, z)$, et cetera, moments can be extracted by applying the moment generating property of PGFs. Indeed, since $U(\lambda, z)$, $Q(\lambda, z)$, et cetera, are analytic at $(\lambda = 0, z = 1)$, the order of taking derivatives can be changed (first to λ and then to z or vice versa).*

Remark 14. *The light-traffic formulas are valid under the assumption that $z^c - T_c(A(\lambda, z))$ is aperiodic (assumption 4 from the introduction). Indeed, the approximation is based on the series expansion of the set of equations (2.14) and (2.15), which only make sense in case of aperiodicity of $z^c - T_c(A(\lambda, z))$.*

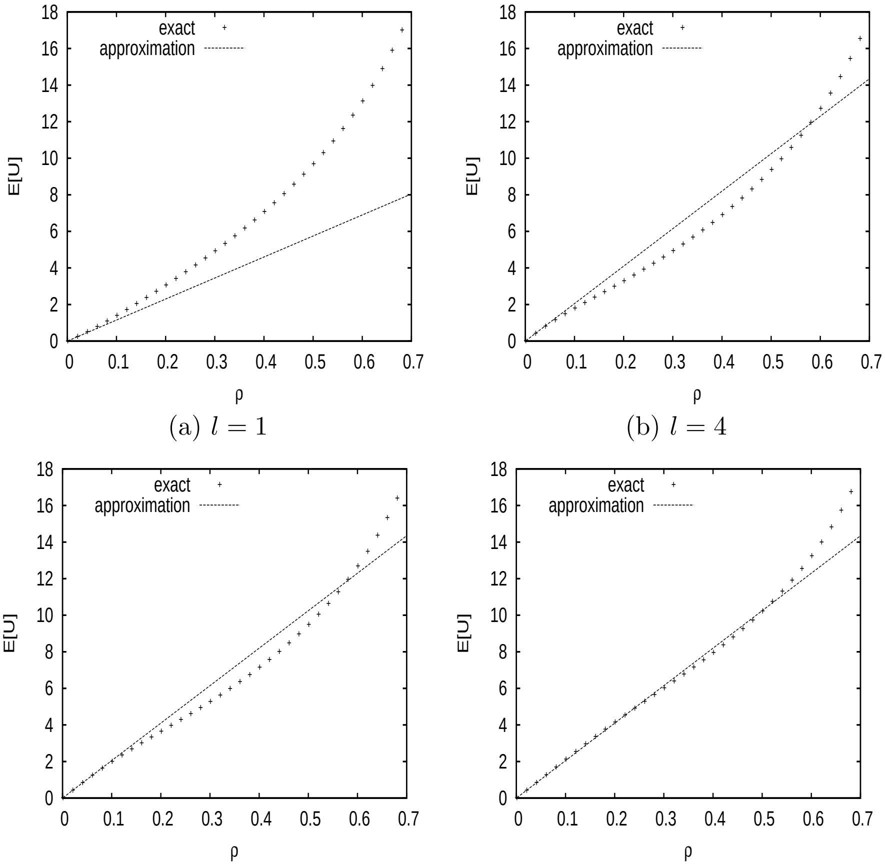
Remark 15. *When $\beta = 0$, our approach leads to full-analytic expressions for $l = 1$ and $l = c$ (the interested reader is referred to our paper [43], where this is demonstrated for the special case whereby $T_n(z) = T_c(z)$ for all n).*

3.2.8 Evaluation of the approximation formulas

In the previous sections, we have deduced light-traffic approximations for a spectrum of quantities related to the buffer content. The purpose of the current section is to assess the accuracy of our approach. Let us therefore consider an example with server capacity 10, a Poisson distribution for the number of arrivals and the service time of a batch of n customers being geometrically distributed with mean value $8 + 0.2n$.

We have depicted in Fig. 3.1-3.4 the light-traffic approximations as well as the exact values of the mean system content and the filling degree (defined as the mean number of customers in a served batch divided by the server capacity c) versus the load ρ, for various values of l and β. We observe that the approximations are accurate in case of light traffic (i.e. small values of ρ, thus small values of λ), and that the larger the service threshold, the longer the range of values of ρ where the approximations remain accurate.

Next, Fig. 3.1 and 3.3 exhibit that when $\beta > 0$, the mean system content as well as the filling degree tend to zero when $\lambda \to 0$ (and thus $\rho \to 0$). This can be explained as follows: when eventually customers arrive for small λ, they will be served after some relatively short time (because $\beta \neq 0$) and this period is negligible as compared to the nearly infinite time until new customers arrive. As a result, the system is virtually always empty. In addition, the nearly infinite interarrival times in combination with $\beta \neq 0$ imply that the system practically always alternates between periods whereby the server is not serving and periods whereby the server processes 0 customers, meaning that the served batches are almost always empty. Finally, we perceive that when $\beta = 0$, the mean system content (Fig. 3.2) and the filling degree (Fig. 3.4) tend to respectively $(l - 1)/2$ and l/c for $\lambda \to 0$. Actually, this is a consequence of the Poisson distribution for the amount of arrivals in a slot. Indeed, when $\lambda \to 0$, the customers are very likely to arrive alone. As a result, the system content nearly always increases by one, until the service threshold l is reached. At that moment, service is initiated with l customers and since the service



(a) $l = 1$

(b) $l = 4$

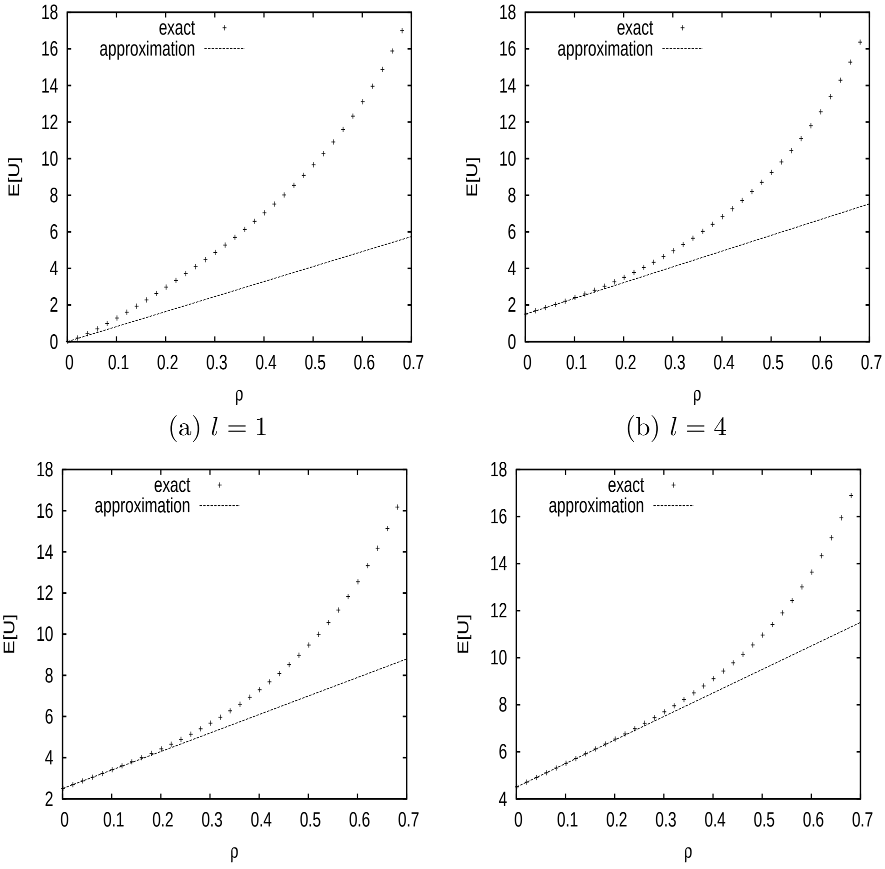
(c) $l = 6$

(d) $l = 10$

Figure 3.1: Evaluation of light-traffic approximation for $\mathrm{E}[U]$ (via formula (3.25)); Poisson arrivals, $\beta = 0.1$, $c = 10$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$

times are negligible as compared to the interarrival times, the system content is uniformly distributed between 0 and $l - 1$, which has mean value $(l - 1)/2$.

Remark 16. *We have also checked the approximations for the queue content, the server content, et cetera, and we have also considered other sets for $A(z)$ and $T_j(z)$. The examples confirm the conclusions that are drawn in this section. We have not depicted all examples in order to keep this dissertation concise.*



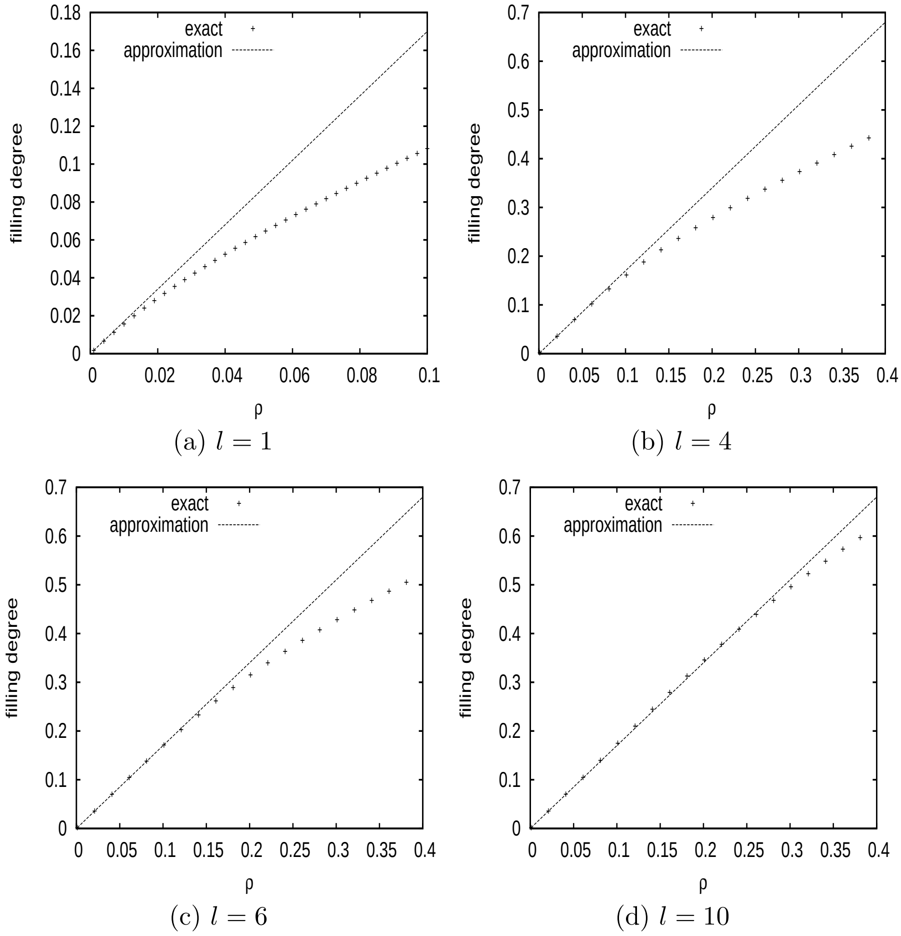
(a) $l = 1$

(b) $l = 4$

(c) $l = 6$

(d) $l = 10$

Figure 3.2: Evaluation of light-traffic approximation for $\mathrm{E}\,[U]$ (via formula (3.31)); Poisson arrivals, $\beta = 0$, $c = 10$, T_n geometrically distributed, $\mathrm{E}\,[T_n] = 8 + 0.2n$



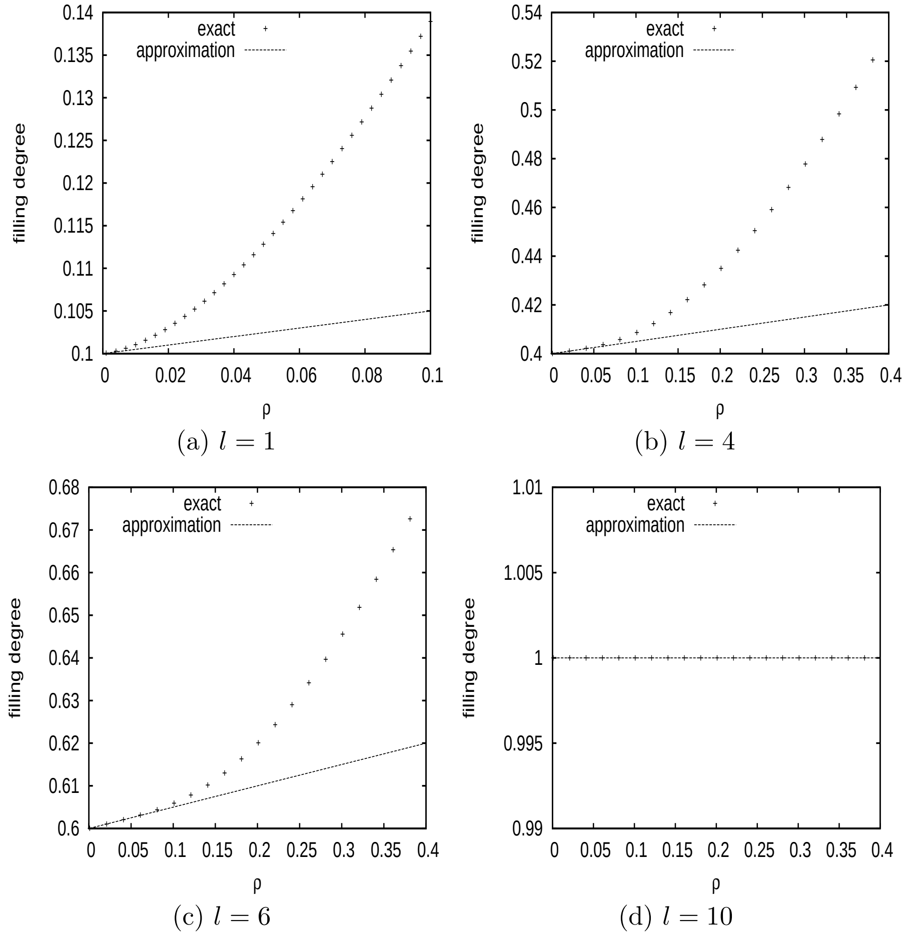
(a) $l = 1$

(b) $l = 4$

(c) $l = 6$

(d) $l = 10$

Figure 3.3: Evaluation of light-traffic approximation for the filling degree (via formula (3.36)); Poisson arrivals, $\beta = 0.1$, $c = 10$, T_n geometrically distributed, $\mathrm{E}\,[T_n] = 8 + 0.2n$



(a) $l = 1$

(b) $l = 4$

(c) $l = 6$

(d) $l = 10$

Figure 3.4: Evaluation of light-traffic approximation for the filling degree (via formula (3.37)); Poisson arrivals, $\beta = 0$, $c = 10$, T_n geometrically distributed, $\mathrm{E}\,[T_n] = 8 + 0.2n$

3.3 Heavy-traffic approximations

In this section, we study the heavy-traffic behaviour of the buffer-related quantities that were computed in chapter 2, by letting λ going to $c/\mathrm{E}\,[T_c]$. We thus analyse the system close to instability ($\rho \to 1$).

3.3.1 System content at random slot boundaries

We here deduce heavy-traffic approximations for $\mathrm{E}\,[U]$, $\mathrm{Var}\,[U]$ and $\Pr\,[U=n]$.

First, let $N_U(z)$ and $D_U(z)$ represent respectively the numerator and denominator of $U(z)$, i.e.

$$\begin{aligned} N_U(z) =&(z^c-1)T_c(A(z))[1-A(z)]\sum_{n=0}^{l-1} d(n)z^n \\ &+\beta\sum_{n=0}^{l-1} d(n)g_n(z)+\sum_{n=l}^{c-1} d(n)h_n(z) \ , \end{aligned} \tag{3.38}$$

$$D_U(z) =[1-A(z)]\,[z^c-T_c(A(z))] \ , \tag{3.39}$$

with $g_n(z)$ and $h_n(z)$ defined as

$$\begin{aligned} g_n(z) \triangleq&(z^n-z^c)T_n(A(z))T_c(A(z))+z^n(z^c-1)T_c(A(z))A(z) \\ &-z^c(z^n-1)T_n(A(z)) \ , \\ h_n(z) \triangleq&T_n(A(z))z^c\{1-z^n-T_c(A(z))\}-T_c(A(z))z^n\{1-z^c-T_n(A(z))\} \ . \end{aligned}$$

It is clear that $N_U(1)=N_U^{'}(1)=D_U(1)=D_U^{'}(1)=0$ and that, owing to the normalisation condition ($U(1)=1$), $N_U^{''}(1)=D_U^{''}(1)$. As a result, we have, after applying l'Hôpital's rule several times:

$$\mathrm{E}\,[U]=U^{'}(1)=\frac{N_U^{'''}(1)-D_U^{'''}(1)}{3D_U^{''}(1)} \ , \tag{3.40}$$

and

$$U^{''}(1)=\frac{3N_U^{''''}(1)D_U^{''}(1)-3D_U^{''''}(1)D_U^{''}(1)-4D_U^{'''}(1)N_U^{'''}(1)+4D_U^{'''}(1)^2}{18D_U^{''}(1)^2} \ . \tag{3.41}$$

Second, we let λ go to $c/\mathrm{E}\,[T_c]$. Then, it is clear that $D_U^{''}(1)=-2\lambda[c-\mathrm{E}\,[T_c]\,\lambda]$ tends to zero. In order to calculate the limit of the numerator and its derivatives at $z=1$, it is required to deduce the limit behaviour of the c unknowns $d(n)$, $n=0,\ldots,c-1$. Recall that these c unknowns can be found by solving the set of c equations (3.1) and (3.2). When λ goes to $c/\mathrm{E}\,[T_c]$, the left-hand-side of (3.2) vanishes. As a result, it can be shown that the solution of the set of equations equals $d(n)=0$. Hence,

$$\lim_{\lambda\uparrow\frac{c}{\mathrm{E}[T_c]}} d(n)=0 \ , \qquad 0\le n\le c-1 \ . \tag{3.42}$$

As a consequence, the numerator of $U(z)$ and all its derivatives become zero at $z=1$.
In conclusion, as $D_U^{''}(z)$ as well as the numerator of $U(z)$ and all its derivatives

become zero at $z = 1$, and on account of (3.40) and (3.41), $U^{'}(1)$ and $U^{''}(1)$ go to infinity according to the following expressions:

$$U^{'}(1) \sim \frac{-D_U^{'''}(1)}{3D_U^{''}(1)} \ ,$$

$$U^{''}(1) \sim \frac{2}{9}\left(\frac{D_U^{'''}(1)}{D_U^{''}(1)}\right)^2 \ .$$

Relying on expression (3.39) for $D_U(z)$ and the moment generating property of PGFs ($\mathrm{E}[U] = U^{'}(1)$ and $\mathrm{Var}[U] = U^{''}(1) - U^{'}(1)^2 + U^{'}(1)$) yields

$$\mathrm{E}[U] \sim \frac{\lambda^3 T_c^{''}(1) + 2\lambda A^{''}(1)\mathrm{E}[T_c] - A^{''}(1)c - \lambda c(c-1)}{2\lambda(c - \mathrm{E}[T_c]\lambda)} \ , \tag{3.43}$$

$$\mathrm{Var}[U] \sim \left[\frac{\lambda^3 T_c^{''}(1) + 2\lambda A^{''}(1)\mathrm{E}[T_c] - A^{''}(1)c - \lambda c(c-1)}{2\lambda(c - \mathrm{E}[T_c]\lambda)}\right]^2 \ . \tag{3.44}$$

Finally, on account of (3.42) and expressions (3.38) and (3.39) for respectively $N_U(z)$ and $D_U(z)$, it follows that $U(z) \to 0$ for $|z| < 1$, which, in turn, implies that $\Pr[U = n] \to 0$ for finite n.

3.3.2 Queue content at random slot boundaries

The heavy-traffic approximation for the queue content at random slot boundaries can be computed completely analogously as the approximation for the system content, and is equal to

$$\mathrm{E}[Q] \sim \frac{\lambda^3 T_c^{''}(1) + 2\lambda A^{''}(1)\mathrm{E}[T_c] - A^{''}(1)c - \lambda c(c-1)}{2\lambda(c - \mathrm{E}[T_c]\lambda)} \ , \tag{3.45}$$

$$\mathrm{Var}[Q] \sim \left[\frac{\lambda^3 T_c^{''}(1) + 2\lambda A^{''}(1)\mathrm{E}[T_c] - A^{''}(1)c - \lambda c(c-1)}{2\lambda(c - \mathrm{E}[T_c]\lambda)}\right]^2 \ , \tag{3.46}$$

and $\Pr[Q = n] \to 0$ for finite n. The formulas are thus exactly equal to those for the system content. This is because the server content becomes negligible as compared to the (very large) queue content.

3.3.3 System content at service completion times

Before we can deduce a heavy-traffic approximation for the system content at service completion times, it is necessary to study the heavy-traffic behaviour of $F(1,1)$ (note that $F(1,1)$ is by definition equal to the probability that the remaining service time equals one, thus the probability that a random slot is the last slot of a service period). Taking the limit $\lambda \to c/\mathrm{E}[T_c]$ in expression (2.19) for $F(1,1)$ would require application of l'Hôpital's rule and consequently the calculation of derivatives of $d(n)$ at $\lambda \to c/\mathrm{E}[T_c]$. Alternatively, we can rely on the rate-in-rate-out principle, which states that when the system is in steady state, the mean number of customers leaving the system per slot equals the mean number of arrivals per slot. The mean number of arrivals equals λ, whereas the mean number of customers leaving the system equals the probability that the service finishes at a random slot ($F(1,1)$) multiplied by the number of customers in service in that slot, which goes to c (this will be proved in the next subsection). Hence, as $\lambda \to c/\mathrm{E}[T_c]$, we find

$$\lim_{\lambda \uparrow \frac{c}{\mathrm{E}[T_c]}} F(1,1) = \frac{1}{\mathrm{E}[T_c]} \quad .$$

Next, let $N_{\tilde{U}}(z)$ and $D_{\tilde{U}}(z)$ represent respectively the numerator and denominator of $\tilde{U}(z)$. It holds that $N_{\tilde{U}}(1) = D_{\tilde{U}}(1) = 0$, and the normalisation condition implies that $N'_{\tilde{U}}(1) = D'_{\tilde{U}}(1)$. As a result, we have, after applying l'Hôpital's rule several times

$$\begin{aligned}
\mathrm{E}\left[\tilde{U}\right] &= \frac{N''_{\tilde{U}}(1) - D''_{\tilde{U}}(1)}{2D'_{\tilde{U}}(1)} \quad , \\
\tilde{U}''(1) &= \frac{2N'''_{\tilde{U}}(1)D'_{\tilde{U}}(1) - 2D'_{\tilde{U}}(1)D'''_{\tilde{U}}(1) - 3D''_{\tilde{U}}(1)N''_{\tilde{U}}(1) + 3D''_{\tilde{U}}(1)^2}{6D'_{\tilde{U}}(1)^2} \quad .
\end{aligned}$$

As $D'_{\tilde{U}}(1)$ (equal to $F(1,1)[c - \mathrm{E}[T_c]\lambda]$) as well as all derivatives of $N_{\tilde{U}}(z)$ at $z = 1$ tend to 0 (because $d(n) \to 0$) when $\lambda \to c/\mathrm{E}[T_c]$, we find that $\mathrm{E}\left[\tilde{U}\right]$ and $\tilde{U}''(1)$ tend to infinity according to the following expressions:

$$\begin{aligned}
\mathrm{E}\left[\tilde{U}\right] &\sim -\frac{D''_{\tilde{U}}(1)}{2D'_{\tilde{U}}(1)} \quad , \\
\tilde{U}''(1) &\sim \frac{1}{2}\left(\frac{D''_{\tilde{U}}(1)}{D'_{\tilde{U}}(1)}\right)^2 \quad .
\end{aligned}$$

Hence,

$$\mathrm{E}\left[\tilde{U}\right] \sim \frac{T''_c(1)\lambda^2 + \mathrm{E}[T_c]A''(1) - c(c-1)}{2[c - \mathrm{E}[T_c]\lambda]} \quad , \tag{3.47}$$

$$\mathrm{Var}\left[\tilde{U}\right] \sim \left(\frac{c(c-1) - T''_c(1)\lambda^2 - \mathrm{E}[T_c]A''(1)}{2[c - \mathrm{E}[T_c]\lambda]}\right)^2 \quad . \tag{3.48}$$

3.3.4 Server content at random slot boundaries

Letting $\lambda \to c/\mathrm{E}[T_c]$ in expression (2.20) for $S(z)$, taking into account that $d(n) \to 0$ and application of l'Hôpital's rule yields

$$\begin{aligned}
S(z) \to &- z^c \frac{c}{\mathrm{E}[T_c]} \sum_{n=0}^{l-1} d'(n) - z^c \beta \sum_{n=0}^{l-1} d'(n) \left\{\mathrm{E}[T_n]\frac{c}{\mathrm{E}[T_c]} - n - \frac{c}{\mathrm{E}[T_c]}\right\} \\
&- z^c \sum_{n=l}^{c-1} d'(n) \left\{\mathrm{E}[T_n]\frac{c}{\mathrm{E}[T_c]} - n\right\} \quad ,
\end{aligned} \tag{3.49}$$

with

$$d'(n) \triangleq \lim_{\lambda \to \frac{c}{\mathrm{E}[T_c]}} \frac{\partial}{\partial \lambda} d(\lambda, n) \quad .$$

Next, as the left-hand-side of (3.2) is not zero for $\lambda < c/\mathrm{E}[T_c]$, not all derivatives of $d(n)$ at $\lambda \to c/\mathrm{E}[T_c]$ ($d'(n)$) are equal to zero. Although $d'(n)$ is difficult to calculate, (3.49) reveals that only terms corresponding to z^c are different from zero. Hence,

$$\lim_{\lambda \to \frac{c}{\mathrm{E}[T_c]}} S(z) = z^c \ ,$$

and consequently

$$\lim_{\lambda \to \frac{c}{\mathrm{E}[T_c]}} \Pr\left[S = n\right] \to \begin{cases} 1 & \text{if } n = c \ , \\ 0 & \text{else } . \end{cases} \tag{3.50}$$

Hence, $\mathrm{E}\left[S\right] \to c$ and $\mathrm{Var}\left[S\right] \to 0$. This result states that the server content is nearly always equal to c, which is no surprise, as we have previously deduced that the queue content goes to infinity if the load tends to one.

3.3.5 Number of customers in a served batch

As $S(z) = z^c$, it follows that $\tilde{S}(z) = z^c$. This can also be observed from expression (2.21) for $\tilde{S}(z)$. Indeed, letting $\lambda \to c/\mathrm{E}\left[T_c\right]$ and appyling l'Hôpital's rule shows that only the terms corresponding to z^c do not vanish. As a result, the corresponding probabilities become

$$\lim_{\lambda \to \frac{c}{\mathrm{E}[T_c]}} \Pr\left[\tilde{S} = n\right] = \begin{cases} 1 & \text{if } n = c \ , \\ 0 & \text{else } . \end{cases} \tag{3.51}$$

3.3.6 Probability that the server processes

The combination of expression (2.22) for Pr [server processes], $q_0(n) = (1 - \beta)d(n)$ for $n = 0, \ldots, l-1$, and $d(n) \to 0$ produces

$$\Pr\left[\text{server processes}\right] \to 1 \ . \tag{3.52}$$

3.3.7 Queue content when the server not processes

As the server nearly always processes when λ approaches $c/\mathrm{E}\left[T_c\right]$, this approximation is not useful. In addition, calculating an approximation would require the calculation of derivatives of $d(n)$ at $\lambda = c/\mathrm{E}\left[T_c\right]$, which is very complicated.

Remark 17. *Expressions (3.43)-(3.52) highlight that the heavy-traffic behaviour is independent of l and β. The reason is that almost always c or more customers are present, so that, regardless of the service policy, a new service is initiated immediately.*

Remark 18. *The appearance of $A^{''}(1)$ and $T_c^{''}(1)$ in the numerator of (3.43)-(3.46) and (3.47)-(3.48) shows that a larger variance in the service times as well as amount of per-slot arrivals, leads to an increasing system content and queue content, which is a typical result in queueing theory. However, although it is also common in queueing theory that $A^{'''}(1)$ and $T_c^{'''}(1)$ appear in the expressions of the variance of the system and queue content, they are not present in the heavy-traffic approximation of* $\mathrm{Var}\left[U\right]$, $\mathrm{Var}\left[Q\right]$ *and* $\mathrm{Var}\left[\tilde{U}\right]$. *The reason is that we only retain the dominant term. For instance, consider expression*

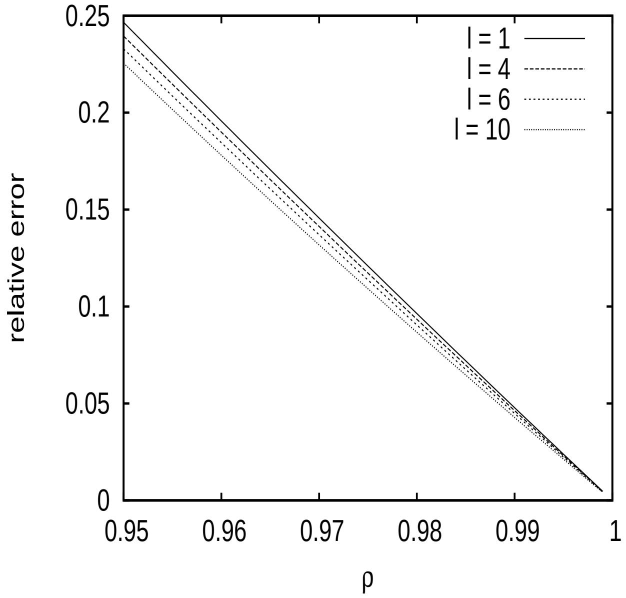
(3.41) for $U^{''}(1)$*. The fourth term in the numerator is dominant and* $A^{'''}(1)$ *and* $T_c^{'''}(1)$ *appear in the other terms. Hence, although* $A^{'''}(1)$ *and* $T_c^{'''}(1)$ *play a role in* $\mathrm{Var}\left[U\right]$*,* $\mathrm{Var}\left[Q\right]$ *and* $\mathrm{Var}\left[\tilde{U}\right]$*, their influence is negligible in case of heavy traffic.*

3.3.8 Evaluation of the approximation formulas

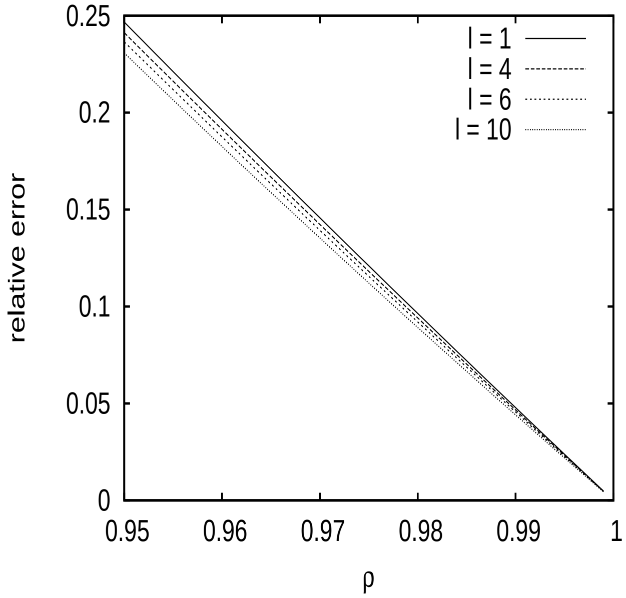
We evaluate approximation formulas (3.43)-(3.52) by considering the same example as in sections 2.5 and 3.2.8 (Poisson arrivals, geometrically distributed service times with mean value dependent on the number of customers in the served batch and a server capacity equal to 10). As the system content goes to infinity (see subsection 3.3.1), we do not plot the approximations and the exact values of $\mathrm{E}\left[U\right]$ as it would lead to a distorted view. Therefore, we study the relative error of $\mathrm{E}\left[U\right]$, defined as follows:

$$\frac{2|\mathrm{E}\left[U\right]-\mathrm{E}\left[U\right]_a|}{\mathrm{E}\left[U\right]+\mathrm{E}\left[U\right]_a},$$

where $\mathrm{E}\left[U\right]_a$ represents the approximation for $\mathrm{E}\left[U\right]$. The relative error of the system content is depicted versus the load ρ in Fig. 3.5, for various values of l and β. Analogously, the relative error of the filling degree is depicted in Fig. 3.6. We observe that the approximations are accurate for a large arrival intensity and that it fits better when l is larger. This is logical, since the approximations exploit the fact that the server nearly always processes at full capacity in case of heavy traffic. We also perceive that the value of β has only a small influence on the relative difference, which can be explained because β is a mechanism that matters especially for small loads.

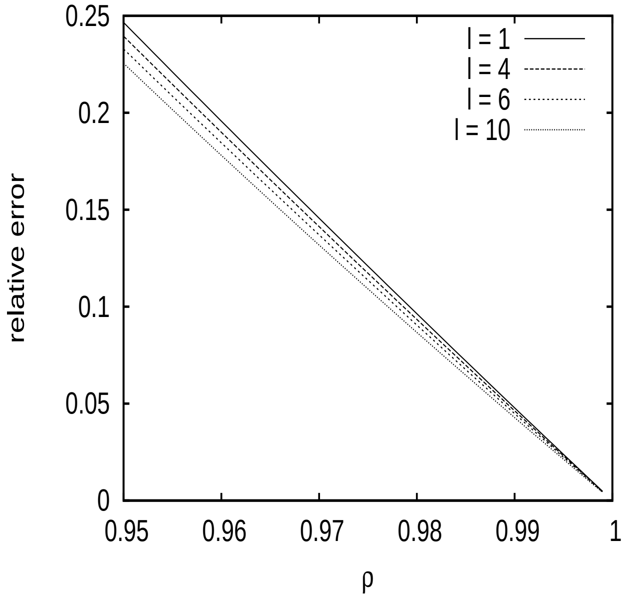


(a) $\beta = 0$

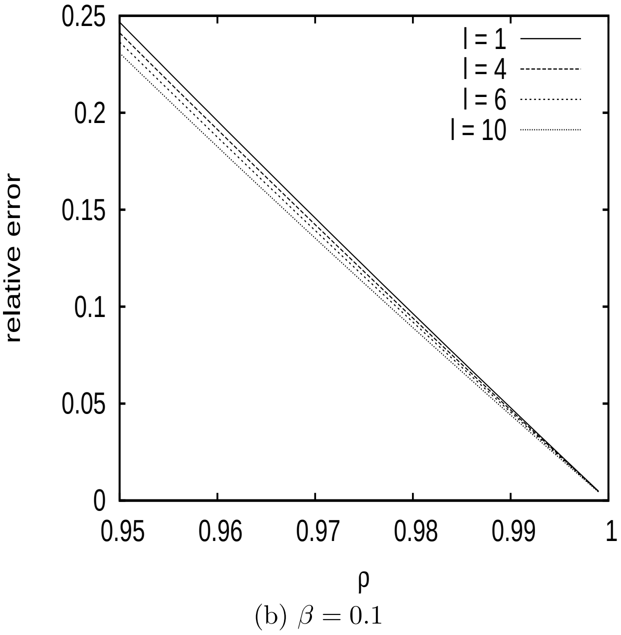


(b) $\beta = 0.1$

Figure 3.5: Relative error between $\mathrm{E}[U]$ and its heavy-traffic approximation; Poisson arrivals, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$



(a) $\beta = 0$



(b) $\beta = 0.1$

Figure 3.6: Relative error between the filling degree and its heavy-traffic approximation; Poisson arrivals, T_n geometrically distributed, $\mathrm{E}\,[T_n] = 8 + 0.2n$

Chapter 4

Customer delay: PGFs and moments

4.1 Preface

Whereas we have studied the buffer content in the preceding chapters, we now turn our focus to the customer delay. As mentioned previously, we define the delay of a customer as the integral number of slots it sojourns in the queue. Hence, the remaining time of the slot wherein the customer arrives as well as the service time is excluded. The structure of this chapter is broadly speaking similar as in chapter 2. We first establish a joint PGF (section 4.2), from which we deduce then various quantities in section 4.3. Next, in section 4.4, we show how performance measures can be extracted from these quantities and finally we demonstrate that these performance measures are useful tools to evaluate real-life batch-service queueing systems. As opposed to chapter 2, only moments can be extracted from the obtained quantities. Calculating tail probabilities, on the other hand, requires another approach, which is the topic of chapter 5.

In our papers [39] and [40], we have analysed the customer delay in a model that is included as a special case in the model discussed throughout this dissertation: it adopts the full-batch service-policy and single-slot service times ($\beta = 0$, $l = c$ and $T_c(z) = z$). In [41], we have studied the customer delay in a model with immediate-batch service-policy ($\beta = 0$, $l = 1$) and a model with full-batch service policy ($\beta = 0$, $l = c$). Both models have in common that geometric service times are considered that are independent of the amount of served customers ($T_n(z) = z/(\mathrm{E}[T_c] - (\mathrm{E}[T_c] - 1)z)$). Finally, in [43], we have extended [39], [40] and [41] by adopting the threshold-based service policy ($\beta = 0$) and considering general service times that are independent of the number of customers being processed. In this chapter, we immediately discuss the versatile model from this dissertation, as it runs along the same lines as in

[39], [40], [41] and [43].

4.2 Joint PGF

The delay of a randomly tagged customer can be subdivided in two parts. The first component (W_1) is the time required to serve batches containing previously arrived customers and is referred to as **queueing delay**. The second part (W_2) is the time needed, starting at the end of the queueing delay, to fill the batch containing the tagged customer with at least l customers or until the server has permission to initiate service with less than l customers and is therefore denominated by **postponing delay**. On account of these definitions it holds that the delay W of a randomly tagged customer equals

$$W = W_1 + W_2 \ .$$

It is important to point out that dependence exists between the queueing delay W_1 and the postponing delay W_2. Indeed, the longer the queueing delay, the more customers arrive during the queueing delay, thus W_1 influences the amount of customers at the beginning of W_2 (this number is denoted by P). If P is larger than or equal to the service threshold l, W_2 equals zero. Otherwise, it may take some time until at least l customers have accumulated or until the server decides to initiate a new service anyway.

In Fig. 4.1, W_1, W_2 and P together with some other notations are illustrated through an example whereby $\beta = 0$, $l = c = 10$ and $T_c(z) = z$. Here, J denotes the tagged customer's arrival slot, Q_J is the queue content at the beginning of this slot and A^- and A^+ represent the amount of arrivals during slot J, respectively before and behind the tagged customer. It is assumed that $Q_J + A^- = 23$ and $A^+ = 2$. As at slot mark $J+1$ the number of customers before the tagged customer equals $Q_J + A^- = 23$, $c = 10$ and $T_c(z) = z$, the queueing delay equals 2 slots. Next, as 3 customers are still in front of the tagged customer after the queueing delay, $A^+ = 2$ and 3 customers have arrived during the queueing delay, 9 customers (the tagged customer included) are in the system after the queueing delay, which is smaller than $l = c = 10$. As a customer arrives after two slots, which thus leads to a sufficient number of present customers, the postponing delay W_2 equals two slots. Hence, the total customer delay W equals $W_1 + W_2 = 2 + 2 = 4$ slots.

The purpose of this section is to compute the joint PGF $\tilde{W}(z,x)$ of the queueing delay W_1 and the postponing delay W_2, i.e.

$$\tilde{W}(z,x) \triangleq \mathrm{E}\left[z^{W_1}x^{W_2}\right] \ .$$

We thereby exploit that W_1 only influences W_2 through the amount of customers at the end of W_1 (P) and thus that W_1 and W_2 are independent if P is given:

$$\begin{aligned}\tilde{W}(z,x) &= \sum_{p=1}^{\infty} \mathrm{Pr}\left[P=p\right] \mathrm{E}\left[z^{W_1}x^{W_2}|P=p\right] \\ &= \sum_{p=1}^{\infty} \mathrm{Pr}\left[P=p\right] \mathrm{E}\left[z^{W_1}|P=p\right] \mathrm{E}\left[x^{W_2}|P=p\right] \ . \qquad (4.1)\end{aligned}$$

Next, we compute $\mathrm{E}\left[x^{W_2}|P=p\right]$. Therefore, we make use of the following relation between W_2 and P:

$$\mathrm{Pr}[W_2 > m|P=p] = \mathrm{Pr}[p + \hat{A}_1 + \cdots + \hat{A}_m < l](1-\beta)^{m+1} \ , \qquad m \geq 0 \ , \qquad (4.2)$$

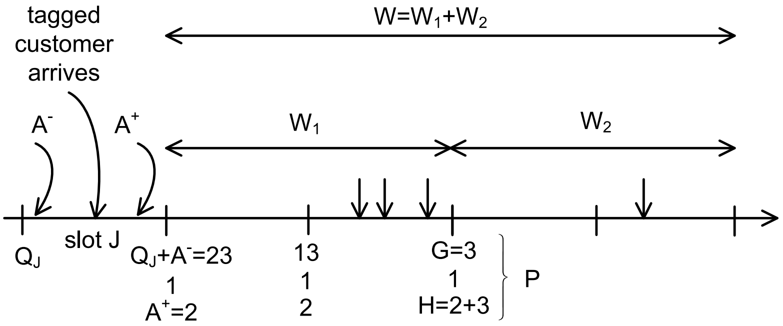


Figure 4.1: Illustration of W_1, W_2 and P and other notations and relations between them; $\beta = 0$, $l = c = 10$, $T_c(z) = z$

with $\hat{A}_j$ the number of arrivals during the j-th slot after the end of the queueing delay. Indeed, W_2 is larger than m if m slots after the queueing delay the number of present customers, $p + \hat{A}_1 + \ldots + \hat{A}_m$, is still smaller than the service threshold l and if the server decides at least during $m + 1$[1] slots not to start a new service with fewer customers. Multiplication of both sides of (4.2) by x^m and summing over all m yields

$$\begin{aligned}
\frac{\mathrm{E}\left[x^{W_2}|P = p\right] - 1}{x - 1} &= \sum_{m=0}^{\infty} x^m \mathrm{Pr}\left[p + \hat{A}_1 + \cdots + \hat{A}_m < l\right] (1-\beta)^{m+1} \\
&= \sum_{m=0}^{\infty} x^m \sum_{n=0}^{l-1} \mathrm{Pr}\left[p + \hat{A}_1 + \cdots + \hat{A}_m = n\right] (1-\beta)^{m+1} \\
&= \sum_{m=0}^{\infty} x^m \sum_{n=0}^{l-1} \frac{1}{n!} \frac{\partial^n}{\partial y^n} y^p A(y)^m \bigg|_{y=0} (1-\beta)^{m+1} \\
&= \sum_{n=p}^{l-1} \frac{(1-\beta)}{n!} \frac{\partial^n}{\partial y^n} \frac{y^p}{1 - (1-\beta)xA(y)} \bigg|_{y=0} ,
\end{aligned}$$

whereby step 3 makes use of the probability generating property of PGFs and the IID character of the arrival process and step 4 takes into account that the n-th derivative is equal to zero for $n < p$. The last equation requires that $|(1 - \beta)xA(y)| < 1$ in the neighbourhood of $y = 0$. We thus have that:

$$\mathrm{E}\left[x^{W_2}|P = p\right] = 1 + (x-1) \sum_{n=p}^{l-1} \frac{(1-\beta)}{n!} \frac{\partial^n}{\partial y^n} \frac{y^p}{1 - (1-\beta)xA(y)} \bigg|_{y=0} . \qquad (4.3)$$

Note that the second term of (4.3) vanishes if $p \geq l$. Indeed, when $p \geq l$, the postponing delay is equal to zero. Substituting (4.3) into (4.1) produces:

$$\tilde{W}(z, x) = P(z, 1) + (x-1) \sum_{n=0}^{l-1} \frac{(1-\beta)}{n!} \frac{\partial^n}{\partial y^n} \frac{P(z, y)}{1 - (1-\beta)xA(y)} \bigg|_{y=0} , \qquad (4.4)$$

[1] Note the appearance of $m+1$ instead of m. In case of m, the server would initiate a new service at the $m+1$-th slot after the queueing delay, so that W_2 would be equal to m which is thus not larger than m.

with

$$P(z,y) \quad \triangleq \quad \mathrm{E}\left[z^{W_1}y^{P}\right] \ .$$

In order to compute $P(z,y)$, we first calculate the joint PGF $\hat{W}(z,x,y)$ of W_1, the number of customers ahead (G) and the number of customers behind (H) the tagged customer at the end of the queueing delay (see Fig. 4.1 for an illustration of these variables), i.e.

$$\hat{W}(z,x,y) \quad \triangleq \quad \mathrm{E}\left[z^{W_1}x^{G}y^{H}\right] \ .$$

Since P is equal to $G+H+1$, $P(z,y)$ is then equal to $y\hat{W}(z,y,y)$. As we consider general service times, we have to bear in mind that a service might be going on during slot J that can continue during slot $J+1$. Therefore, we consider two situations depending on whether the remaining service time at the beginning of slot J, R_J, equals 0 or not:

- $R_J = 0$. In this case, the server is not processing during slot J. As a consequence, the server can start a new service at slot $J+1$ if there are enough customers or if the server is allowed to initiate a new service with fewer customers (with probability β). As the number of previously arrived customers equals $Q_J + A^-$ and the server can process c customers simultaneously, the queueing delay W_1 equals $\left\lfloor \frac{Q_J+A^-}{c} \right\rfloor$ service periods[2], whereby each service time is distributed according to $T_c(z)$. After the queueing delay, $(Q_J + A^-)$ mod c previously arrived customers are still in front of the tagged customer[3]. Hence, $G = (Q_J + A^-)$ mod c. These customers are served in the same batch as the tagged customer. Finally, the amount of customers behind the tagged customer at the end of the queueing delay is the sum of the number of customers that arrive after the tagged one during slot J and during the queueing delay. Hence, $H = A^+ + \sum_{i=1}^{W_1} A_{J+i}$.

- $R_J \geq 1$. In this case, the server first continues $R_J - 1$ slots with the current service period. After that, $Q_J + A^-$ customers are ahead of the tagged one and another $\left\lfloor \frac{Q_J+A^-}{c} \right\rfloor$ service periods are part of W_1 (each length is distributed according to $T_c(z)$). Hence, $W_1 = \left\lfloor \frac{Q_J+A^-}{c} \right\rfloor$ service periods $+R_J - 1$. Analogously as in the first case, $G = (Q_J + A^-)$mod c and $H = A^+ + \sum_{i=1}^{W_1} A_{J+i}$.

We split the computation of the joint PGF $\hat{W}(z,x,y)$ of W_1, G and H in two parts corresponding to these two situations:

$$\hat{W}(z,x,y) \quad = \quad \mathrm{E}\left[z^{W_1}x^{G}y^{H}\{R_J = 0\}\right] + \mathrm{E}\left[z^{W_1}x^{G}y^{H}\{R_J \geq 1\}\right] \ . \qquad (4.5)$$

[2] We adopt the standard convention that $\lfloor . \rfloor$ denotes the floor function, i.e. $\lfloor x \rfloor \triangleq \max\{n \in \mathbb{Z} : n \leq x\}$.

[3] "mod" is the common notation for the modulo operator.

For the first component we have:

$$\mathrm{E}\left[z^{W_1}x^G y^H\{R_J=0\}\right]=\sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}d(nc+m,k,0)T_c(zA(y))^n x^m y^k\ , \tag{4.6}$$

with

$$d(n,m,k)\triangleq\Pr\left[Q_J+A^-=n,A^+=m,R_J=k\right]\ .$$

Due to the IID nature of the customer arrivals, we can write the corresponding PGF $D(z,x,y)$ as the following product:

$$\begin{aligned}D(z,x,y)\triangleq&\mathrm{E}\left[z^{Q_J+A^-}x^{A^+}y^{R_J}\right]\\=&\mathrm{E}\left[z^{A^-}x^{A^+}\right]V(z,1,y)\ ,\end{aligned} \tag{4.7}$$

with $V(z,x,y)$ the joint PGF of the queue content, the server content and the remaining service time that was computed in chapter 2:

$$\begin{aligned}V(z,x,y)\left[1-\frac{A(z)}{y}\right]=&(1-\beta)\left[1-\frac{A(z)}{y}\right]\sum_{n=0}^{l-1}d(n)z^n\\&+\left(\frac{x}{z}\right)^c T_c(y)[A(z)-1]\sum_{n=0}^{l-1}d(n)z^n\\&+\beta\sum_{n=0}^{l-1}d(n)\left[x^nT_n(y)-z^n\left(\frac{x}{z}\right)^cT_c(y)A(z)\right]\\&+\left(\frac{x}{z}\right)^cT_c(y)A(z)F(z,1)-A(z)F(z,x)\\&+\sum_{n=l}^{c-1}d(n)\left[x^nT_n(y)-z^n\left(\frac{x}{z}\right)^cT_c(y)\right]\ ,\end{aligned} \tag{4.8}$$

with

$$\begin{aligned}&z^cA(z)F(z,x)\left[z^c-T_c(A(z))\right]\\&\quad=z^cx^cT_c(A(z))[A(z)-1]\sum_{n=0}^{l-1}d(n)z^n\\&\qquad+\beta x^cT_c(A(z))\sum_{n=0}^{l-1}d(n)\left[z^cT_n(A(z))-z^nT_c(A(z))A(z)\right]\\&\qquad+x^cT_c(A(z))\sum_{n=l}^{c-1}d(n)\left[z^cT_n(A(z))-z^nT_c(A(z))\right]\\&\qquad+\beta[z^c-T_c(A(z))]\sum_{n=0}^{l-1}d(n)\left[z^cx^nT_n(A(z))-x^cz^nT_c(A(z))A(z)\right]\\&\qquad+[z^c-T_c(A(z))]\sum_{n=l}^{c-1}d(n)\left[z^cx^nT_n(A(z))-x^cz^nT_c(A(z))\right]\ ,\end{aligned} \tag{4.9}$$

and whereby the unknowns $d(n)$ can be calculated by solving the set of equations (2.14)-(2.15). In [24], it is proved, by taking into account that an arbitrary customer is more likely to arrive in a slot with more customer arrivals, that $\mathrm{E}\left[z^{A^-}x^{A^+}\right]$ is equal to

$$\mathrm{E}\left[z^{A^-}x^{A^+}\right]=\frac{A(z)-A(x)}{\lambda(z-x)}\ . \tag{4.10}$$

In order to relate $\mathrm{E}\left[z^{W_1}x^Gy^H\{R_J=0\}\right]$ with $D(z,x,y)$, we first introduce some notations. The function $u(z,y)$ is defined as the "principal c-th root" of $T_c(zA(y))$, i.e.

$$u(z,y) \triangleq |T_c(zA(y))|^{1/c}e^{\imath \mathrm{Arg}(T_c(zA(y)))/c} \ , \tag{4.11}$$

with $\imath$ the imaginary unit and $\mathrm{Arg}(z)$ the principal value of the argument of z, i.e. a mapping in the interval $]-\pi,\pi]$. Next, $\delta\langle l=j\rangle$ is the Kronecker-Delta function (i.e. $\delta\langle l=j\rangle=1$ if $l=j$ and $\delta\langle l=j\rangle=0$ if $l\neq j$). We now rewrite expression (4.6) for $\mathrm{E}\left[z^{W_1}x^Gy^H\{R_J=0\}\right]$ as follows:

$$\begin{aligned}&\mathrm{E}\left[z^{W_1}x^Gy^H\{R_J=0\}\right]\\ &\quad=\sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1}d(nc+m,k,0)u(z,y)^{nc+m-j}x^jy^k\delta\langle m=j\rangle \ .\end{aligned} \tag{4.12}$$

On account of the standard property

$$\delta\langle m=j\rangle=\sum_{i=0}^{c-1}\frac{1}{c}\varepsilon_i^{nc+m-j} \ ,$$

with ε_i the i-th complex c-th root of 1 ($\varepsilon_i \triangleq e^{\imath 2\pi i/c}$), (4.12) can further be transformed into

$$\begin{aligned}&\mathrm{E}\left[z^{W_1}x^Gy^H\{R_J=0\}\right]\\ &\quad=\sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1}d(nc+m,k,0)u(z,y)^{nc+m-j}x^jy^k\sum_{i=0}^{c-1}\frac{1}{c}\varepsilon_i^{nc+m-j}\\ &\quad=\frac{1}{c}\sum_{i=0}^{c-1}D(u(z,y)\varepsilon_i,y,0)\sum_{j=0}^{c-1}u(z,y)^{-j}x^j\varepsilon_i^{-j}\\ &\quad=\frac{u(z,y)^c-x^c}{cu(z,y)^c}\sum_{i=0}^{c-1}D(u(z,y)\varepsilon_i,y,0)\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i-x} \ .\end{aligned} \tag{4.13}$$

We continue with the second term of (4.5). In a similar way as formula (4.13), we find

$$\begin{aligned}\mathrm{E}\left[z^{W_1}x^Gy^H\{R_J\geq 1\}\right]&=\frac{1}{czA(y)}\frac{u(z,y)^c-x^c}{u(z,y)^c}\\ &\cdot\sum_{i=0}^{c-1}\left[D(u(z,y)\varepsilon_i,y,zA(y))-D(u(z,y)\varepsilon_i,y,0)\right]\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i-x} \ .\end{aligned} \tag{4.14}$$

Substitution of (4.13) and (4.14) in (4.5) produces:

$$\begin{aligned}\hat{W}(z,x,y)=&\frac{u(z,y)^c-x^c}{cu(z,y)^czA(y)}\\ &\cdot\left[[zA(y)-1]\sum_{i=0}^{c-1}D(u(z,y)\varepsilon_i,y,0)\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i-x}\right.\\ &\left.+\sum_{i=0}^{c-1}D(u(z,y)\varepsilon_i,y,zA(y))\frac{u(z,y)\varepsilon_i}{u(z,y)\varepsilon_i-x}\right] \ .\end{aligned} \tag{4.15}$$

Making use of formulas (4.7), (4.10), (4.11) and expressions (4.8) and (4.9) for $V(z,x,y)$ and $F(z,x)$ yields:

$$
\begin{aligned}
\hat{W}(z,x,y) = & \frac{T_c(zA(y)) - x^c}{c\lambda T_c(zA(y))} \sum_{i=0}^{c-1} \frac{[A(u(z,y)\varepsilon_i) - A(y)]u(z,y)\varepsilon_i}{[u(z,y)\varepsilon_i - y][u(z,y)\varepsilon_i - x][zA(y) - A(u(z,y)\varepsilon_i)]} \\
& \cdot \left\{ [zA(y)-1]\,(1-\beta) \sum_{n=0}^{l-1} d(n)(u(z,y)\varepsilon_i)^n + \sum_{n=l}^{c-1} d(n)\,[T_n(zA(y)) - (u(z,y)\varepsilon_i)^n] \,\}\right. \\
& \left. + \, \beta \sum_{n=0}^{l-1} d(n)\,[T_n(zA(y)) - (u(z,y)\varepsilon_i)^n] \right\} \quad .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P(z,y) \quad &= \quad \mathrm{E}\left[z^{W_1} y^P\right] \\
&= \quad y\hat{W}(z,y,y) \\
&= \quad y \frac{T_c(zA(y)) - y^c}{c\lambda T_c(zA(y))} \sum_{i=0}^{c-1} \frac{A(u(z,y)\varepsilon_i) - A(y)}{[u(z,y)\varepsilon_i - y]^2} \frac{u(z,y)\varepsilon_i}{zA(y) - A(u(z,y)\varepsilon_i)} \\
& \qquad \cdot \left\{ [zA(y)-1]\,(1-\beta) \sum_{n=0}^{l-1} d(n)(u(z,y)\varepsilon_i)^n \right. \\
& \qquad + \beta \sum_{n=0}^{l-1} d(n)\,[T_n(zA(y)) - (u(z,y)\varepsilon_i)^n] \\
& \qquad \left. + \sum_{n=l}^{c-1} d(n)\,[T_n(zA(y)) - (u(z,y)\varepsilon_i)^n] \right\} \quad . \qquad (4.16)
\end{aligned}
$$

Substitution of (4.11) and (4.16) in (4.4) produces the final expression for the joint PGF $\tilde{W}(z,x)$. As was the case in chapter 2, the obtained joint PGF enables us to extract several quantities. In the next section, we derive the PGF of the total customer delay and the marginal PGFs of the queueing and the postponing delay.

Remark 19. *In case of single (Bernoulli) arrivals, the analysis is simplified considerably. Indeed, in that case, we find, by substituting $A(y)$ by $1-\lambda+\lambda y$ in (4.3) and applying Leibniz's rule for the derivative of a product that*

$$
\mathrm{E}\left[x^{W_2}|P=p\right] = 1 + (1-\beta)(x-1)\frac{1 - \left[\frac{(1-\beta)x\lambda}{1-(1-\beta)x(1-\lambda)}\right]^{l-p}}{1-(1-\beta)x} \quad , \qquad p < l \ ,
$$

so that no derivatives have to be calculated anymore. When $\beta = 0$, W_2 given $P = p$ becomes a sum of $l-p$ geometrically distributed random variables. Indeed, the time until an arrival is then geometrically distributed with parameter λ and $l-p$ arrivals are required to initiate service.

Remark 20. *Note the appearance of $u(z,y)^c$ in the denominator of the right-hand side of (4.15), which suggests that zeroes of $u(z,y)^c$ (among which $z = 0, \forall y$ because $T_c(0) = 0$) might lead to poles of $\hat{W}(z,x,y)$. However (4.15) can* *be transformed into*

$$\hat{W}(z,x,y)u(z,y)^c = \frac{x^c}{czA(y)} \sum_{i=0}^{c-1}\sum_{j=1}^{c}\left(\frac{u(z,y)\varepsilon_i}{x}\right)^j \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{k=0}^{\infty} d(n,m,k)(u(z,y)\varepsilon_i)^n y^m \left[[zA(y)-1]\delta\langle k=0\rangle + [zA(y)]^k\right] \ .$$

Since

$$\sum_{i=0}^{c-1} \varepsilon_i^{j+n} = \begin{cases} c & \textit{if } j+n \textit{ is a multiple of } c \ , \\ 0 & \textit{else} \ , \end{cases}$$

and $j \geq 1$, it is clear that $u(z,y)^c$ is also a factor of the numerator of (4.15). Hence, zeroes of $u(z,y)^c$ are no poles of $\hat{W}(z,x,y)$.

4.3 Quantities related to the customer delay

4.3.1 Customer delay

Since the customer delay W is the sum of the queueing and the postponing delay, $W(z)$ is found by substituting x by z in expression (4.4) for the joint PGF of W_1 and W_2:

$$W(z) = \tilde{W}(z,z) = P(z,1) + (z-1)\sum_{n=0}^{l-1} \frac{(1-\beta)}{n!}\frac{\partial^n}{\partial y^n}\left.\frac{P(z,y)}{1-(1-\beta)zA(y)}\right|_{y=0} \ . \qquad (4.17)$$

4.3.2 Queueing delay

The PGF $W_1(z)$ of the queueing delay is found by summing out W_2 in $\tilde{W}(z,x)$. Hence, substituting x by 1 in (4.4) gives:

$$W_1(z) \triangleq \mathrm{E}\left[z^{W_1}\right] = \tilde{W}(z,1) = P(z,1) \ . \qquad (4.18)$$

4.3.3 Postponing delay

The PGF $W_2(z)$ of the postponing delay is established by summing out W_1 in $\tilde{W}(z,x)$. Hence, substituting z by 1 and x by z in (4.4) produces:

$$\begin{aligned} W_2(z) &\triangleq \mathrm{E}\left[z^{W_2}\right] \\ &= \tilde{W}(1,z) \\ &= 1 + (z-1)\sum_{n=0}^{l-1}\frac{(1-\beta)}{n!}\frac{\partial^n}{\partial y^n}\left.\frac{P(1,y)}{1-(1-\beta)zA(y)}\right|_{y=0} \ . \end{aligned} \qquad (4.19)$$

4.4 Performance measures

As expression (4.17) for $W(z)$ contains derivatives, it is quite hard to extract performance measures from it. Especially the calculation of tail probabilities seems impossible. Therefore, we consider a different approach for obtaining tail probabilities in chapter 5. Moments, however, can be computed from (4.17).

As the mean delay can also be obtained from Little's law [57], [79] ($\mathrm{E}[W] = \mathrm{E}[Q]/\lambda$), we demonstrate in this section how the variance of the delay can be deduced from (4.17). From (4.17), we find that the second derivative of $W(z)$ evaluated at $z = 1$ reads

$$W^{''}(1) = W_1^{''}(1) + 2(1-\beta)\sum_{n=0}^{l-1}\frac{1}{n!}\left[\frac{\partial^n}{\partial y^n}\left.\frac{P^{(1)}(1,y)}{1-(1-\beta)A(y)}\right|_{y=0} + \frac{\partial^n}{\partial y^n}\left.\frac{P(1,y)(1-\beta)A(y)}{[1-(1-\beta)A(y)]^2}\right|_{y=0}\right] ,$$

with

$$P^{(1)}(1,y) \triangleq \left.\frac{\partial}{\partial z}P(z,y)\right|_{z=1} .$$

We can calculate $\frac{1}{n!}\frac{\partial^n}{\partial y^n}\left.\frac{P^{(1)}(1,y)}{1-(1-\beta)A(y)}\right|_{y=0}$ and $\frac{1}{n!}\frac{\partial^n}{\partial y^n}\left.\frac{P(1,y)(1-\beta)A(y)}{[1-(1-\beta)A(y)]^2}\right|_{y=0}$ by inverting respectively $\frac{P^{(1)}(1,y)}{1-(1-\beta)A(y)}$ and $\frac{P(1,y)(1-\beta)A(y)}{[1-(1-\beta)A(y)]^2}$ (for instance with the inverse discrete Fast Fourier Transform).
Finally, on account of the moment generating property of PGFs, the variance of W reads

$$\mathrm{Var}[W] = W^{''}(1) - \mathrm{E}[W]^2 + \mathrm{E}[W] .$$

In addition, moments of the queueing delay can be obtained by applying the moment generating property of PGFs to (4.18) and moments for the postponing delay can be extracted from (4.19) along the same lines as we deduced $\mathrm{Var}[W]$ from (4.17).

4.5 Numerical examples

In this section, we demonstrate that moments of the customer, queueing and postponing delay are useful tools to evaluate batch-service queueing systems. Let us therefore again consider the example with a Poisson distribution for the number of customer arrivals in a random slot, a server capacity equal to 10 and geometrically distributed service times with mean value dependent on the number of customers n in the served batch: $\mathrm{E}[T_n] = 8 + 0.2n$.

In Fig. 4.2, $\mathrm{E}[W]$, $\mathrm{Var}[W]$, $\mathrm{E}[W_1]$ and $\mathrm{E}[W_2]$ are depicted versus the load ρ for the case $\beta = 0$. We perceive that the customer delay goes to infinity when the load tends to one and that it is caused by the queueing delay (W_1). Next, we observe that if $l = 1$, the customer delay tends to zero for $\rho \to 0$. Indeed, when in this case a customer eventually arrives, the system is almost certainly empty, so that the customer is served the next slot. On the other hand, when $l > 1$, the customer delay goes to infinity if $\rho \to 0$ and this effect is caused by the postponing delay (W_2). Fig. 4.2 further exhibits that a larger service threshold l has an advantageous impact on the queueing delay and a negative effect on the postponing delay. Indeed, a larger threshold leads to a smaller probability that a service just has started when a new customer arrives and implies that the already present customers have to wait longer until enough customers are present. As the postponing delay dominates in case of light traffic, it is better to adopt a small threshold in that case and since for heavy traffic the queueing delay is dominant, larger thresholds become steadily preferable when the load increases.

Whereas the influence of l is studied in Fig. 4.2, Fig. 4.3 evaluates the influence of β. We perceive that when $\rho \to 0$, the customer delay does not go to infinity anymore if $\beta \neq 0$. In the sequel, we intuitively calculate $\mathrm{E}[W]$, $\mathrm{E}[W_1]$ and $\mathrm{E}[W_2]$ in this case. When $\lambda \to 0$, the interarrival times go to infinity. As a result, the system nearly always alternates between periods whereby the server is not processing ("idle period", with mean length $(1-\beta)/\beta$) and periods whereby the server processes zero customers ("active period", with mean length $\mathrm{E}[T_0]$). If, eventually, a customer arrives in an active period, it will probably be the only one in system at service completion so that it suffers on average a postponing delay of $(1-\beta)/\beta$ slots. When the customer arrives during an idle period, $\mathrm{E}[W_2]$ is, owing to the memoryless property of the geometric distribution, also equal to $(1-\beta)/\beta$. Hence,

$$\mathrm{E}[W_2] \to \frac{1-\beta}{\beta} \quad .$$

Next, the queueing delay W_1 of a customer equals k if it arrives in an active period that lasts another k slots after the arrival slot. The probability that a customer arrives in an active period of length n becomes

$$\begin{aligned}\mathrm{Pr}[\text{arrive in active period of length n}] \to &\mathrm{Pr}[\text{arrive in active period}]\\ &.\ \mathrm{Pr}[\text{length active period equals n}]\\ =&\frac{\mathrm{E}[T_0]}{\mathrm{E}[T_0]+(1-\beta)/\beta}\frac{n\mathrm{Pr}[T_0=n]}{\mathrm{E}[T_0]}\\ =&\frac{n\mathrm{Pr}[T_0=n]}{\mathrm{E}[T_0]+(1-\beta)/\beta} \quad ,\end{aligned}$$

whereby we have taken into account that a customer is more likely to arrive in a long active period ([24]). Hence, since the position of the customer's arrival slot in the active period is uniformly distributed, we have

$$\mathrm{Pr}[W_1=k] \to \sum_{n=k+1}^{\infty} \frac{\mathrm{Pr}[T_0=n]}{\mathrm{E}[T_0]+(1-\beta)/\beta} \quad .$$

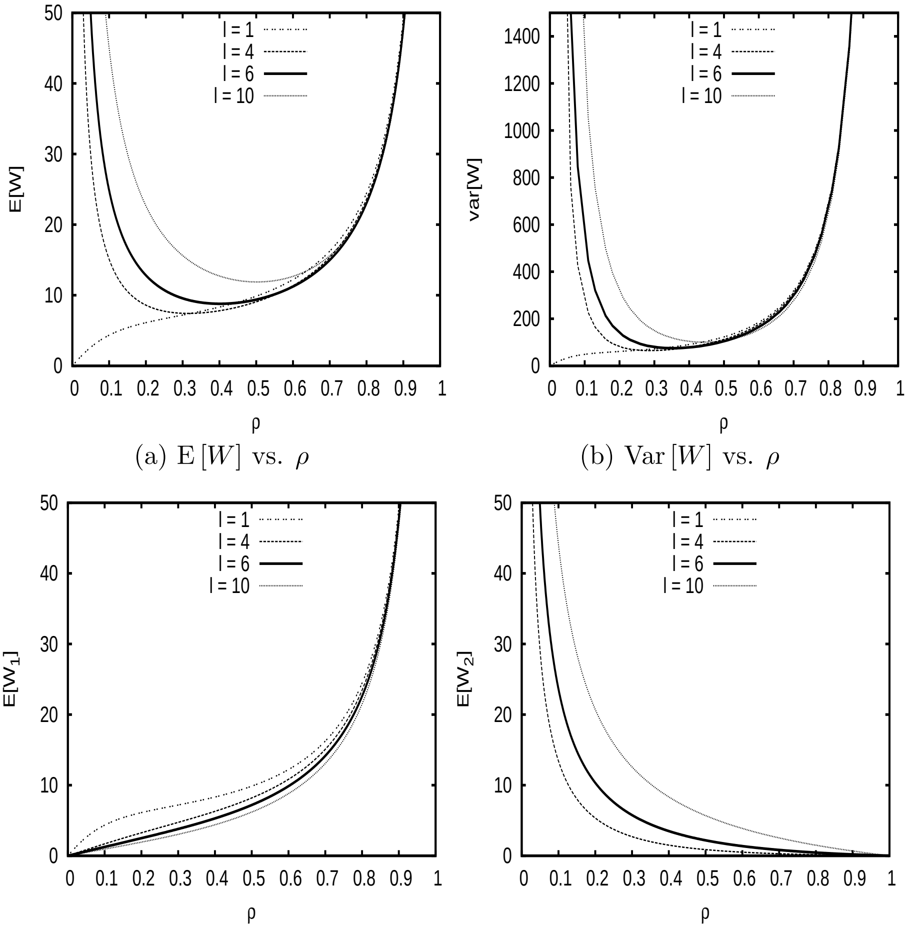
As a result,

$$\begin{aligned}\mathrm{E}[W_1] \to &\sum_{k=1}^{\infty} k \sum_{n=k+1}^{\infty} \frac{\mathrm{Pr}[T_0=n]}{\mathrm{E}[T_0]+(1-\beta)/\beta}\\ =&\frac{1}{2}\frac{\mathrm{Var}[T_0]+\mathrm{E}[T_0]^2-\mathrm{E}[T_0]}{\mathrm{E}[T_0]+(1-\beta)/\beta} \quad ,\end{aligned}$$

and consequently, as $W = W_1 + W_2$,

$$\mathrm{E}[W] \to \frac{1-\beta}{\beta}+\frac{1}{2}\frac{\mathrm{Var}[T_0]+\mathrm{E}[T_0]^2-\mathrm{E}[T_0]}{\mathrm{E}[T_0]+(1-\beta)/\beta} \quad .$$

Finally, Fig. 4.3 also reveals that $\beta = 0$ is the best option for heavy traffic, but the impact is negligible as then nearly always l or more customers are present at service completion times.



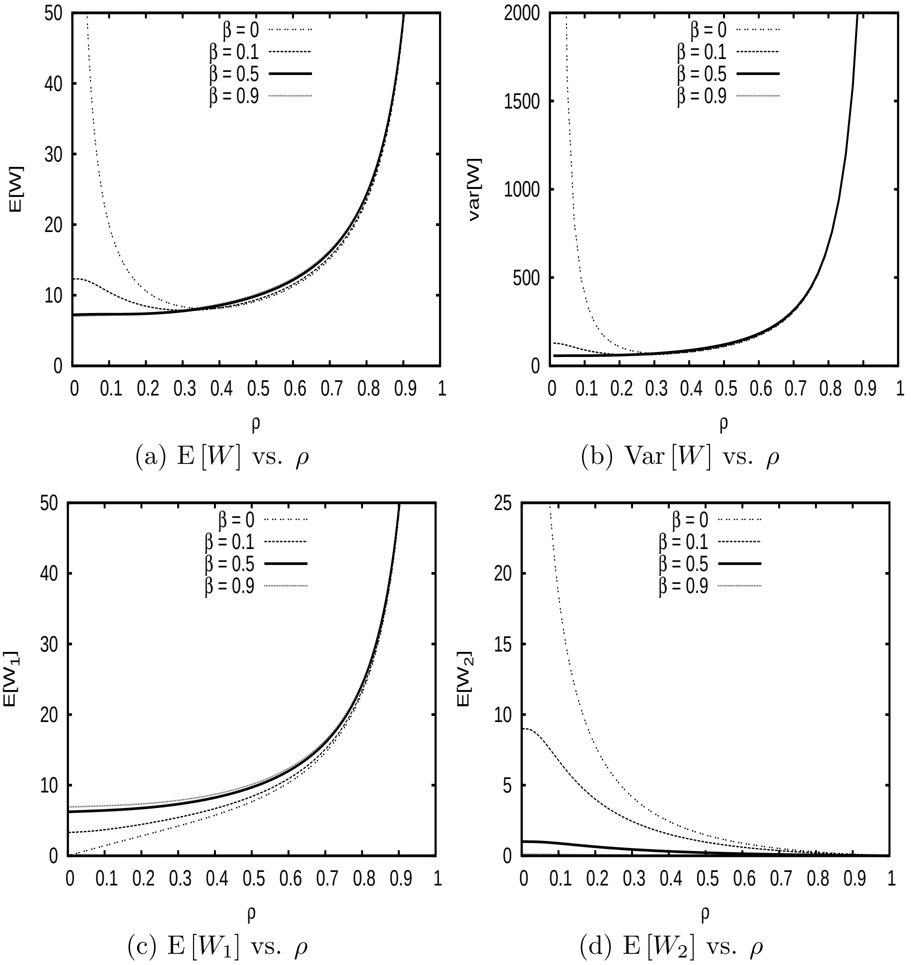
(a) $\mathrm{E}[W]$ vs. ρ

(b) $\mathrm{Var}[W]$ vs. ρ

(c) $\mathrm{E}[W_1]$ vs. ρ

(d) $\mathrm{E}[W_2]$ vs. ρ

Figure 4.2: Influence of service threshold l on the delay; Poisson arrivals, $\beta = 0$, $c = 10$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$



(a) $\mathrm{E}[W]$ vs. ρ

(b) $\mathrm{Var}[W]$ vs. ρ

(c) $\mathrm{E}[W_1]$ vs. ρ

(d) $\mathrm{E}[W_2]$ vs. ρ

Figure 4.3: Influence of β on the delay; Poisson arrivals, $c = 10$, $l = 5$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$

Chapter 5

Customer delay: tail probabilities

5.1 Preface

Tail probabilities of the customer delay form an important performance measure. Consider for instance an edge router where packets with the same QoS and destination are aggregated into bursts, which are then injected into the network. On account of the competition for the transmission channel, packets might suffer a delay. When packets are delay-sensitive (for instance voice packets), they become worthless if their delay becomes large, so they are dropped. The quality of the upperlayer application is then typically expressed in terms of the (order of magnitude of the) probability of this event. Also, in the example of blood pooling from chapter 2, tail probabilities of the delay have to be investigated due to the perishable nature of blood samples.

As formula (4.17) for the customer delay W contains derivatives, it seems impossible to extract tail probabilities from it. Therefore, we follow another approach in this chapter so that tail probabilities can be calculated. In our paper [40], we have deduced tail probabilities in a batch-service queueing model with single-slot service times and full-batch service policy ($\beta = 0$, $l = c$, $T_c(z) = z$). In [44], we have carried out the analysis in a threshold-based model with general service times that are independent of the number of served customers ($\beta = 0$, $T_n(z) = T_c(z)$). Whereas in previous chapters, analysis of a basic model and the final model runs along the same lines, extending a basic model creates several novel difficulties in the context of tail probabilities of the customer delay. Therefore, we first discuss the models from [40] and [44] before dealing with the versatile model from this dissertation. More specifically, we study in section 5.2 the "basic model" from [40]. Then, in section 5.3, we analyse the "intermediate model" from [44] and finally, the "final model" from this dissertation is dealt with in section 5.4.

5.2 Basic model

In this section, we deduce tail probabilities for the model from [40] with single-slot service times and full-batch service policy ($\beta = 0$, $l = c$, $T_c(z) = z$). Recall that in chapter 4 we have divided the delay of a randomly tagged customer in two components: the queueing delay W_1 and the postponing delay W_2, so that

$$W = W_1 + W_2 \ .$$

The queueing delay is the time to process batches with previously arrived customers and the postponing delay is the time period, starting after the queueing delay, until the batch with the tagged customer constains c customers. The key idea to calculate tail probabilities is to **redefine W_2, so that it starts at the same moment as** W_1. As a result, the delay then becomes a maximum of two parts:

$$W = \max(W_1, \tilde{W}_2) \ , \tag{5.1}$$

with $\tilde{W}_2$ the redefined postponing delay. This is illustrated in Fig. 5.1.

On account of (5.1), we obtain

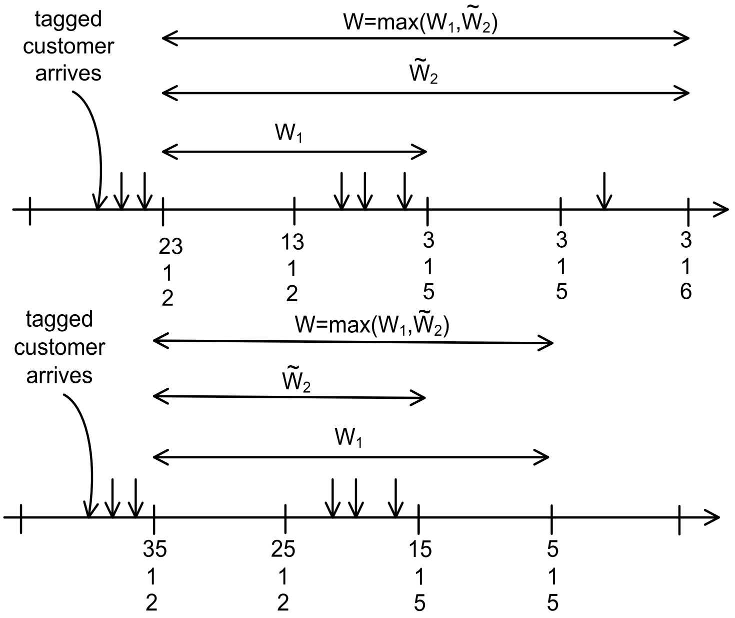


Figure 5.1: Two examples of the new definition of the postponing delay ($c = 10$); three numbers are mentioned at every slot mark, which together represent the system content; they respectively characterise the number of customers before the tagged customer, the tagged customer itself (1) and the number of customers after the tagged customer

$$\begin{aligned}\Pr[W > w] &= \Pr\left[W_1 > w \vee \tilde{W}_2 > w\right] \\ &= \Pr[W_1 > w] + \Pr\left[\tilde{W}_2 > w\right] - \Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right] \quad .\end{aligned}$$

Calculation of joint probabilities of W_1 and $\tilde{W}_2$ is difficult. Therefore, we propose some lower and upper bounds, that only require calculation of marginal tail probabilities of W_1 and $\tilde{W}_2$. The following property paves the path towards establishment of a lower bound for $\Pr[W > w]$:

$$\Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right] \leq \min\left(\Pr[W_1 > w], \Pr\left[\tilde{W}_2 > w\right]\right) \quad . \tag{5.2}$$

A lower bound is obtained by assuming that the equality in (5.2) holds, leading to

$$\Pr[W > w] \geq \max\left(\Pr[W_1 > w], \Pr\left[\tilde{W}_2 > w\right]\right) \quad . \tag{5.3}$$

An upper bound is established from the inequality $\Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right] \geq 0$, leading to:

$$\Pr[W > w] \leq \Pr[W_1 > w] + \Pr\left[\tilde{W}_2 > w\right] \quad . \tag{5.4}$$

These bounds require the calculations of $\Pr[W_1 > w]$ and $\Pr\left[\tilde{W}_2 > w\right]$, which are discussed in the two following subsections respectively.

5.2.1 Calculation of $\Pr[W_1 > w]$

The PGF $W_1(z)$ of the queueing delay in the versatile model was established in chapter 4 (formula (4.18)). The PGF for the basic model is found by setting $\beta = 0$, $l = c$ and $T_c(z) = z$ in (4.18), leading to:

$$W_1(z) = \frac{(z-1)^2}{c\lambda z} \sum_{i=0}^{c-1} \frac{A\left(z^{1/c}\varepsilon_i\right) - 1}{\left(z^{1/c}\varepsilon_i - 1\right)^2} \frac{z^{1/c}\varepsilon_i}{z - A\left(z^{1/c}\varepsilon_i\right)} \sum_{n=0}^{c-1} d(n)\left(z^{1/c}\varepsilon_i\right)^n \quad , \tag{5.5}$$

with $z^{1/c}$ the principal c-th root of z, i.e. $z^{1/c} \triangleq |z|^{1/c} e^{\imath \mathrm{Arg}(z)/c}$, whereby $\imath$ characterises the imaginary unit, $|z|$ is the absolute value of z and $\mathrm{Arg}(z)$ represents the principal value of the argument of z (i.e. it is a mapping in the interval $]-\pi, \pi]$). In addition, ε_i, $0 \leq i \leq c-1$, is the i-th complex c-th root of 1, i.e. $\varepsilon_i \triangleq e^{(\imath 2\pi i)/c}$ and $d(n)$, $0 \leq n \leq c-1$ are unknowns that have to be calculated by solving a set of linear equations (equations (2.14)-(2.15)). Locating the dominant singularity(ies) (i.e the singularity(ies) with smallest modulus) in (5.5) is not easy, and therefore we take a look at $W_1(z^c)$:

$$W_1(z^c) = \frac{(z^c-1)^2 z}{c\lambda z^c} \sum_{i=0}^{c-1} \frac{A(z\varepsilon_i) - 1}{(z\varepsilon_i - 1)^2} \frac{\varepsilon_i}{z^c - A(z\varepsilon_i)} \sum_{n=0}^{c-1} d(n)(z\varepsilon_i)^n \quad .$$

In [106], it is proved that, under the assumptions mentioned in the introduction, $z^c - A(z)$ has a unique zero with smallest modulus larger than one and that it is a real number in $]1, \Re_A[$ with multiplicity one. Let $\hat{z}$ represent that zero. It follows that $\hat{z}\varepsilon_i^{-1}$ is the unique zero with smallest modulus larger than one of $z^c - A(z\varepsilon_i)$. Indeed, if z^* would be a zero of this function with modulus larger than one and smaller than or equal to $|\hat{z}\varepsilon_i^{-1}|$, $z^*\varepsilon_i$ would be a zero of $z^c - A(z)$

and $|z^*\varepsilon_i| = |z^*||\varepsilon_i| = |z^*| \leq |\hat{z}\varepsilon_i^{-1}| = |\hat{z}|$, which is impossible as $\hat{z}$ is the only zero of $z^c - A(z)$ with smallest modulus larger than one. In addition, $\hat{z}\varepsilon_i^{-1}$ is a zero with multiplicity one, because otherwise

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(z^c - A(z\varepsilon_i)\right)\bigg|_{z=\hat{z}\varepsilon_i^{-1}} = c\hat{z}^{c-1}\varepsilon_i - A'(\hat{z})\varepsilon_i = 0 \ ,$$

which would imply that

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(z^c - A(z)\right)\bigg|_{z=\hat{z}} = 0 \ ,$$

meaning that $\hat{z}$ would be a zero of multiplicity larger than one of $z^c - A(z)$. Summarized, $W_1(z^c)$ has c dominant singularities $\hat{z}\varepsilon_i^{-1}$ $(0 \leq i \leq c-1)$ and it are poles with multiplicity one. Hence, $W_1(z^c)$ is in a neighborhood of $\hat{z}\varepsilon_i^{-1}$ proportional to

$$W_1(z^c) \sim \left(1 - \frac{z}{\hat{z}\varepsilon_i^{-1}}\right)^{-1} G_i(z) \ , \tag{5.6}$$

with

$$G_i(z) = \frac{1 - \frac{z}{\hat{z}\varepsilon_i^{-1}}}{z^c - A(z\varepsilon_i)} \frac{(z^c-1)^2 z}{c\lambda z^c} \frac{A(z\varepsilon_i) - 1}{(z\varepsilon_i - 1)^2} \varepsilon_i \sum_{n=0}^{c-1} d(n)(z\varepsilon_i)^n \ ,$$

analytic in the neighbourhood of $\hat{z}\varepsilon_i^{-1}$. As we have located the singularities of $W_1(z^c)$ instead of $W_1(z)$, we have to be careful with the application of Darboux's theorem. First, note that

$$\begin{aligned}
\frac{W_1(z^c) - 1}{z^c - 1} &= \frac{\sum_{n=0}^{\infty} \Pr\left[W_1 = n\right] z^{nc} - 1}{z^c - 1} \\
&= \sum_{n=0}^{\infty} \Pr\left[W_1 = n\right] \frac{(z^c)^n - 1}{z^c - 1} \\
&= \sum_{n=0}^{\infty} \Pr\left[W_1 = n\right] \sum_{w=0}^{n-1} (z^c)^w \\
&= \sum_{w=0}^{\infty} z^{wc} \sum_{n=w+1}^{\infty} \Pr\left[W_1 = n\right] \\
&= \sum_{w=0}^{\infty} \Pr\left[W_1 > w\right] z^{wc} \ .
\end{aligned}$$

In other words, $\Pr[W_1 > w]$ is now the coefficient corresponding to z^{wc} in $[W_1(z^c) - 1]/(z^c - 1)$ instead of the coefficient of z^w in $[W_1(z) - 1]/(z - 1)$. Next, (5.6) implies that

$$\frac{W_1(z^c) - 1}{z^c - 1} \sim \left(1 - \frac{z}{\hat{z}\varepsilon_i^{-1}}\right)^{-1} \frac{G_i(z)}{z^c - 1} \ . \tag{5.7}$$

Application of formula (1.1) of Darboux's theorem on (5.7) and taking into account that $A(\hat{z}) = \hat{z}^c$ then yields

$$\begin{aligned}
\Pr\left[W_1 > w\right] &\approx \sum_{i=0}^{c-1} \frac{G_i(\hat{z}\varepsilon_i^{-1})}{(\hat{z}\varepsilon_i^{-1})^c - 1} (\hat{z}\varepsilon_i^{-1})^{-wc} \\
&= \frac{-\hat{z}^{-(w+1)c}(\hat{z}^c - 1)^2 \sum_{n=0}^{c-1} d(n)\hat{z}^n}{\lambda(\hat{z}-1)^2\left[c\hat{z}^{c-1} - A'(\hat{z})\right]} \ .
\end{aligned} \tag{5.8}$$

5.2.2 Calculation of $\Pr\left[\tilde{W}_2 > w\right]$

Let us first recall some definitions from chapter 4. The tagged customer's arrival slot is designated by J and A^- and A^+ represent the number of customer arrivals during slot J respectively before and after the tagged customer has arrived. As mentioned in chapter 4, $(Q_J + A^-)$ mod c of the previously arrived customers are served in the same batch as the tagged customer. If the sum of this number and the number of arrivals after the tagged customer in slot J as well as in w consecutive slots following slot J, plus one (to take into account the presence of the tagged customer itself), is still less than c, then the postponing delay continues and will thus exceed w. This leads to the following relation:

$$\Pr\left[\tilde{W}_2 > w\right] = \Pr\left[([Q_J + A^-] \bmod c) + 1 + A^+ + \sum_{i=1}^{w} A_{J+i} < c\right] \ . \tag{5.9}$$

In order to compute the right-hand-side of (5.9), we make use of the probability generating property of PGFs. To this end, we commence with the calculation of $\mathrm{E}\left[x^{([Q_J+A^-] \bmod c)}x^{A^+}\right]$:

$$\begin{aligned}
&\mathrm{E}\left[x^{([Q_J+A^-] \bmod c)}x^{A^+}\right] \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty} \Pr\left[Q_J + A^- = nc + m, A^+ = k\right] x^m x^k \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1} \Pr\left[Q_J + A^- = nc + m, A^+ = k\right] x^j x^k \delta\langle m = j\rangle \\
&= \sum_{n=0}^{\infty}\sum_{m=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1} \Pr\left[Q_J + A^- = nc + m, A^+ = k\right] x^j x^k \sum_{i=0}^{c-1} \frac{1}{c}\varepsilon_i^{nc+m-j} \ .
\end{aligned} \tag{5.10}$$

Owing to the IID nature of the customer arrivals and since an arbitrary customer is more likely to arrive in a slot with more customer arrivals [24], the joint PGF of $Q_J + A^-$ and A^+ equals

$$\mathrm{E}\left[z^{Q_J+A^-}x^{A^+}\right] = Q_J(z)\frac{A(z) - A(x)}{\lambda(z - x)} \ .$$

As a result, (5.10) transforms into:

$$\begin{aligned}
\mathrm{E}\left[x^{([Q_J+A^-] \bmod c)}x^{A^+}\right] &= \frac{1}{c}\sum_{i=0}^{c-1}\sum_{j=0}^{c-1}\left(\frac{x}{\varepsilon_i}\right)^j Q_J(\varepsilon_i)\frac{A(\varepsilon_i) - A(x)}{\lambda(\varepsilon_i - x)} \\
&= \frac{1}{c}\sum_{i=0}^{c-1} Q_J(\varepsilon_i)\frac{A(\varepsilon_i) - A(x)}{\lambda(\varepsilon_i - x)}\varepsilon_i\frac{x^c - 1}{x - \varepsilon_i} \ .
\end{aligned}$$

On account of the IID arrival process, Q_J is identically distributed as Q, the queue content at a random slot boundary, whose PGF is equal to, by substituting β by 0, l by c and $T_c(z)$ by z in (2.17)

$$Q(z) = \frac{(z^c - 1)\sum_{n=0}^{c-1} d(n)z^n}{z^c - A(z)} \ .$$

This implies, together with the aperiodicity of $z^c - T_c(A(z))$ (assumption 4 from section 1.6), that $Q(\varepsilon_i) = 0$ for $i = 1, \ldots, c-1$. In addition, the normalization condition of PGFs produces $Q(\varepsilon_0) = Q(1) = 1$. As a consequence, we have

$$\mathrm{E}\left[x^{(Q_J+A^-) \bmod c}x^{A^+}\right] = \frac{A(x) - 1}{\lambda(x - 1)}\frac{x^c - 1}{c(x - 1)} \ . \tag{5.11}$$

Remark 21. *The calculations that lead to formula (5.11) show that $(Q_J + A^-)$ mod c has a uniform distribution between 0 and $c-1$. Indeed, $(Q_J + A^-)$ mod c can be conceived as the serial number of a randomly tagged customer in its batch (starting to count from 0) and due to the full-batch service policy, a batch in service always consists of c customers.*

Application of (5.11) and the probability generating property of PGFs in (5.9) yields

$$\Pr\left[\tilde{W}_2 > w\right] = \sum_{m=0}^{c-1} \frac{1}{m!}\, \frac{\mathrm{d}^m}{\mathrm{d}x^m} xA(x)^w \frac{A(x)-1}{\lambda(x-1)}\, \frac{x^c-1}{c(x-1)}\bigg|_{x=0} .$$

After some mathematical manipulations, this can be transformed into

$$\begin{aligned}\Pr\left[\tilde{W}_2 > w\right] &= \sum_{m=0}^{c-2} \frac{1}{m!}\, \frac{\mathrm{d}^m}{\mathrm{d}x^m} A(x)^w \frac{A(x)-1}{\lambda(x-1)}\, \frac{x^c-1}{c(x-1)}\bigg|_{x=0} \\ &= \frac{1}{c} \sum_{m=0}^{c-2} \sum_{k=0}^{m} \frac{1}{k!(m-k)!}\, \frac{\mathrm{d}^k}{\mathrm{d}x^k}\, \frac{x^c-1}{x-1}\bigg|_{x=0} \frac{\mathrm{d}^{m-k}}{\mathrm{d}x^{m-k}} A(x)^w \frac{A(x)-1}{\lambda(x-1)}\bigg|_{x=0} .\end{aligned}$$

Invoking

$$\frac{x^c-1}{x-1} = \sum_{n=0}^{c-1} x^n ,$$

yields

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k}\, \frac{x^c-1}{x-1}\bigg|_{x=0} = k! , \qquad k < c .$$

Hence

$$\begin{aligned}\Pr\left[\tilde{W}_2 > w\right] &= \frac{1}{c} \sum_{m=0}^{c-2} \sum_{k=0}^{m} \frac{1}{(m-k)!}\, \frac{\mathrm{d}^{m-k}}{\mathrm{d}x^{m-k}} A(x)^w \frac{A(x)-1}{\lambda(x-1)}\bigg|_{x=0} \\ &= \frac{1}{c} \sum_{m=0}^{c-2} \frac{c-1-m}{m!}\, \frac{\mathrm{d}^m}{\mathrm{d}x^m} A(x)^w \frac{A(x)-1}{\lambda(x-1)}\bigg|_{x=0} .\end{aligned} \tag{5.12}$$

Formula (5.12) can be implemented in a mathematical program such as matlab. This procedure suffers from the drawback that high-order derivatives may have to be computed, which causes a considerable reduction in speed and even is infeasible if c is quite large. Therefore, we now deduce an approximation for $\Pr\left[\tilde{W}_2 > w\right]$, whereby no derivatives have to be computed.

Multiplying both sides of (5.12) by z^w , relying on $\sum_{w=0}^{\infty} \Pr\left[\tilde{W}_2 > w\right] z^w = [\tilde{W}_2(z) - 1]/(z-1)$ and taking the sum over all values of w produces:

$$\tilde{W}_2(z) = 1 + \frac{z-1}{c} \sum_{m=0}^{c-2} \frac{c-1-m}{m!}\, \frac{\partial^m}{\partial x^m} \frac{A(x)-1}{\lambda(x-1)}\, \frac{1}{1-zA(x)}\bigg|_{x=0} . \tag{5.13}$$

The m-th ($m \geq 0$) derivative with respect to x of $[A(x)-1]/[\lambda(x-1)](1/[1-zA(x)])$ can be written as

$$\frac{\partial^m}{\partial x^m} \frac{A(x)-1}{\lambda(x-1)}\, \frac{1}{1-zA(x)} = \sum_{j=0}^{m} \frac{C_{m,j}(z,x)}{[1-zA(x)]^{j+1}} , \tag{5.14}$$

whereby $C_{m,j}(z,x)$ are functions of z and x that have no factor $1 - zA(x)$ in their denominator. As opposed to $C_{m,j}(z,x)$ for $j \neq m$, $C_{m,m}(z,x)$ is relatively easy to calculate:

$$C_{m,m}(z,x) = \frac{A(x)-1}{\lambda(x-1)} m! z^m A^{'}(x)^m \quad .$$

The substitution of (5.14) in (5.13) yields

$$\tilde{W}_2(z) = 1 + \frac{z-1}{c} \sum_{m=0}^{c-2} \frac{c-1-m}{m!} \sum_{j=0}^{m} \frac{C_{m,j}(z,0)}{[1-zA(0)]^{j+1}} \quad . \tag{5.15}$$

From this equation, it is clear that $z = 1/A(0)$ is the dominant pole of $\tilde{W}_2(z)$ and that it has multiplicity $c-1$. Consequently, if we retain the simple most dominant term from this expression, which is the one for which $1/A(0)$ has the highest multiplicity, we find that $\tilde{W}_2(z)$ is proportional to

$$\tilde{W}_2(z) \sim \frac{z-1}{c} \frac{1-A(0)}{\lambda} (zA^{'}(0))^{c-2} \left[1 - \frac{z}{1/A(0)}\right]^{-(c-1)} \quad ,$$

in a neighborhood of $z = 1/A(0)$. As $1/A(0)$ is a pole with multiplicity $c-1$, the approximation for $\Pr\left[\tilde{W}_2 > w\right]$ is found by applying formula (1.3) from Darboux's theorem, leading to:

$$\Pr\left[\tilde{W}_2 > w\right] \approx w^{c-2} \frac{1-A(0)}{\lambda} \frac{A(0)^w}{c(c-2)!} \left(\frac{A^{'}(0)}{A(0)}\right)^{c-2} \quad .$$

However, we notice that the value of the approximation for $\Pr\left[\tilde{W}_2 > w\right]$ increases as w increases, for $0 \leq w \leq (2-c)/\ln(A(0))$. When, for instance $c = 10$ and $A(0) = e^{-0.5}$, $(2-c)/\ln(A(0))$ equals 16, which indicates that the approximation is probably inaccurate for w between 0 and 16 (and even for larger w-values) as $\Pr\left[\tilde{W}_2 > w\right]$ is obviously a monotonically decreasing function. We therefore propose a more accurate approximation formula. Notice that we only retained the term with $j = m = c-2$ around $z = 1/A(0)$ in (5.15), as it produces the largest power of $1 - zA(0)$ in the denominator. Instead of only retaining this term, we take all the terms into account for which $j = m$. We thus retain for every m the term that produces the largest power of $1 - zA(0)$ in the denominator. We thus take advantage of the fact that we can easily calculate $C_{m,m}(z,x)$ for all m. Hence, $[\tilde{W}_2(z) - 1]/(z-1)$ transforms in a neighborhood of $z = 1/A(0)$ into

$$\frac{\tilde{W}_2(z)-1}{z-1} \sim \frac{1}{c} \sum_{m=0}^{c-2} \frac{(c-1-m)z^m A^{'}(0)^m}{[1-zA(0)]^{m+1}} \frac{1-A(0)}{\lambda} \quad . \tag{5.16}$$

Next, $1/[1-zA(0)]^{m+1}$ can be rewritten as follows:

$$\begin{aligned}
\frac{1}{[1-zA(0)]^{m+1}} &= \frac{1}{m!A(0)^m} \frac{\mathrm{d}^m}{\mathrm{d}z^m} \frac{1}{1-A(0)z} \\
&= \frac{1}{m!A(0)^m} \frac{\mathrm{d}^m}{\mathrm{d}z^m} \sum_{w=0}^{\infty} [A(0)z]^w \\
&= \frac{1}{m!A(0)^m} \sum_{w=m}^{\infty} A(0)^w z^{w-m} \frac{w!}{(w-m)!} \quad . \qquad (5.17)
\end{aligned}$$

The second step requires that $|A(0)z| < 1$, which is satisfied for z approaching $1/A(0)$ from the left. The substitution of (5.17) in (5.16) produces:

$$\begin{aligned}\frac{\tilde{W}_2(z)-1}{z-1} &\sim \frac{1-A(0)}{c\lambda}\sum_{m=0}^{c-2} A^{'}(0)^m(c-1-m)\sum_{w=m}^{\infty} z^w \frac{w!}{m!(w-m)!}A(0)^{w-m} \\ &= \frac{1-A(0)}{c\lambda}\sum_{w=0}^{\infty} z^w \sum_{m=0}^{\min(c-2,w)} A^{'}(0)^m(c-1-m)\binom{w}{m}A(0)^{w-m} \ .\end{aligned}$$

Equating powers of z^w at both sides of the equation and taking into account that $[\tilde{W}_2(z)-1]/(z-1) = \sum_{w=0}^{\infty}\Pr\left[\tilde{W}_2 > w\right] z^w$, yields

$$\Pr\left[\tilde{W}_2 > w\right] \approx \frac{1-A(0)}{c\lambda}\sum_{m=0}^{\min(c-2,w)} A^{'}(0)^m(c-1-m)\binom{w}{m}A(0)^{w-m} \ . \tag{5.18}$$

Note that for large w, formula (5.18) becomes a sum from 0 to $c-2$. We further point out that the binomial coefficient causes no difficulties, since efficient routines exist to calculate them, even for large w.

Remark 22. *Note that this approach is not suited for cases whereby $A^{'}(0) = 0$, as only the term corresponding to $m = 0$ in (5.16) differs from 0. In these cases, additional terms with $j < m$ must be taken into account in (5.15).*

5.2.3 Evaluation of approximation formulas

In this section, we evaluate the accuracy of our approach. First, we study formula (5.8) for $\Pr[W_1 > w]$. Then, we focus on approximation (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ and finally the accuracy of the lower and upper bounds (respectively formulas (5.3) and (5.4)) for $\Pr[W > w]$ is covered.

In Figures 5.2-5.3, approximation (5.8) as well as simulated values[1] for $\Pr[W_1 > w]$ are depicted versus w for various combinations of server capacities, loads and following distributions for the number of customer arrivals: Poisson ($A(z) = e^{\lambda(z-1)}$), Geometric ($A(z) = 1/(1+\lambda-\lambda z)$ and C-center ($A(z) = 1-\lambda/c+\lambda/(2c)(z^{c-1}+z^{c+1})$). We observe that approximation formula (5.8) is accurate, even for relatively small values of w. The figures further exhibit that higher loads lead to larger tail probabilities of the queueing delay W_1 as expected. In addition, the distribution of the number of per-slot arrivals has an undeniable impact. We notice that the larger the variance in the arrival process (c-center 3.15, Poisson 4.5, geometric 24.75), the slower the probabilities decay, which is a classic result in queueing theory. Finally, we perceive that although the load remains equal (and thus the mean arrival rate λ increases) a larger

[1] Only for $\Pr\left[\tilde{W}_2 > w\right]$ we have an exact formula at our disposal. For the other tail probabilities ($\Pr[W_1 > w]$ and $\Pr[W > w]$) throughout this section, we have therefore depicted the 95% confidence intervals resulting from 10 Monte Carlo simulations whereby each simulation generates W_1 and W for 10^8 customers.

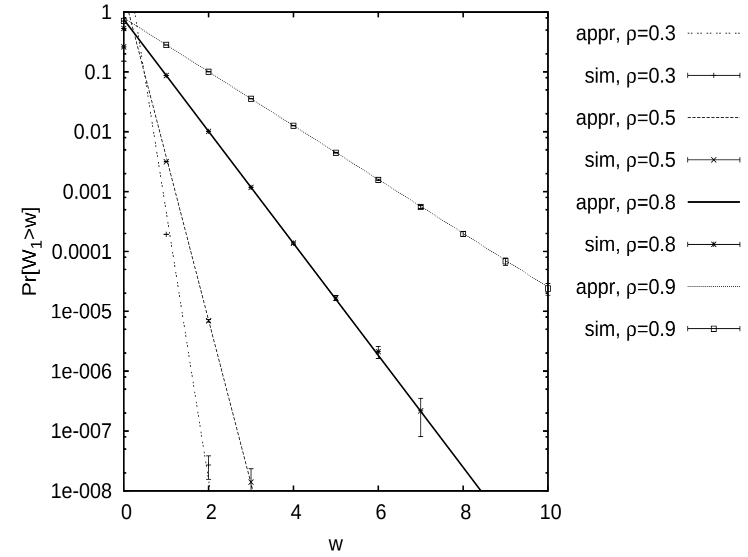
server capacity c has a positive influence on the tail probabilities.

In Figures 5.4-5.5, approximation (5.18) as well as exact values (via formula (5.12)) for $\Pr\left[\tilde{W}_2 > w\right]$ are depicted versus w for various combinations of distributions for the number of customer arrivals, server capacities and loads. We observe that the approximation is inaccurate for small values of w and becomes better for larger values of w. We also perceive that formula (5.18) is not accurate for c-centered arrivals, which is a result of $A^{'}(0)$ being zero in that case (in remark 22, we have mentioned that formula (5.18) cannot be applied when $A^{'}(0) = 0$). Although formula (5.18) is not always very accurate, it might still be useful: in practice, one often only requires the knowledge of the order of magnitude of $\Pr\left[\tilde{W}_2 > w\right]$, which is approximated well by (5.18). In addition, formula (5.18) is a lot faster than exact formula (5.12) that requires the calculation of high-order derivatives. Ultimately, for increasing values of c, the calculation of (5.12) becomes unfeasible, which highlights the necessity of the approximation formula.
Figures 5.4-5.5 further exhibit that a larger load leads to faster decaying tail probabilities of the postponing delay $\tilde{W}_2$. Indeed, when customers arrive more frequently, it takes less time to wait until enough customers have arrived to initiate a new service with c customers.
Next, we observe that a larger variance in the arrivals (Poisson 1.5, geometric 3.75, c-centered 5.55) leads to slower decaying $\Pr\left[\tilde{W}_2 > w\right]$.
Finally, Fig. 5.5 (c) shows that $\Pr\left[\tilde{W}_2 > w\right]$ decreases faster for a larger value of the server capacity c, which seems counterintuitive at first sight, because more customers have to be present before a new service can be initiated. We however have to bear in mind that a larger c also infers a larger mean arrival rate λ, in order to keep the load ρ constant. Hence, two opposite effects occur: a negative from a larger server capacity (more customers have to be present before a new service can be initiated) and a positive from a larger mean arrival rate (customers arrive more frequently). From approximation formula (5.18), it can be seen that the decay rate of $\log(\Pr\left[\tilde{W}_2 > w\right])$ is dominated by $A(0)$, which decreases when λ increases in case of Poisson arrivals, which explains why $\Pr\left[\tilde{W}_2 > w\right]$ decreases faster for a larger server capacity c.

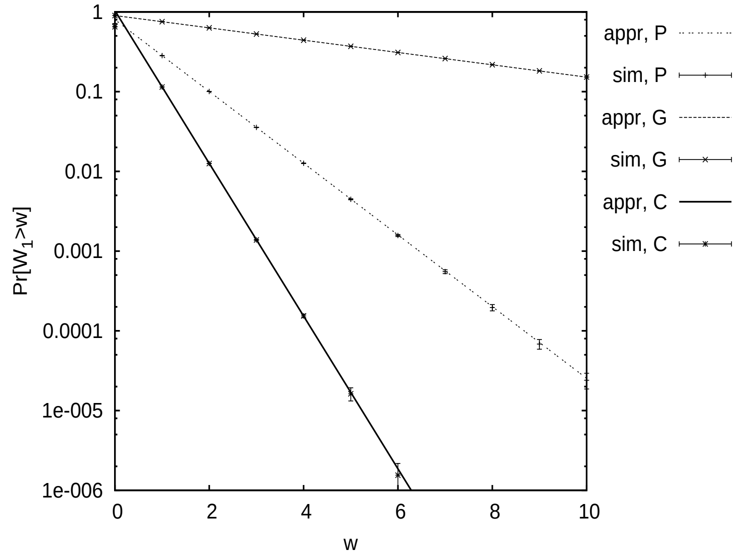
In order to evaluate formulas (5.3) and (5.4) for respectively the lower and upper bound for $\Pr[W > w]$, these are depicted in Figures 5.6-5.7 versus w, together with simulated values. We observe that although the bounds nearly coincide, the bounds are not accurate for small values of w and become better for larger values. We illustrate in Figures 5.8-5.9 that this is mainly caused by the approximation for $\Pr\left[\tilde{W}_2 > w\right]$: if the exact values of $\Pr\left[\tilde{W}_2 > w\right]$ are used, the bounds become extremely accurate. As approximations for $\Pr\left[\tilde{W}_2 > w\right]$ are accurate for large values of w, it then becomes possible to calculate the bounds by making use of the approximation for $\Pr\left[\tilde{W}_2 > w\right]$. Unfortunately, we cannot show $\Pr[W > w]$ for larger values of w in Fig. 5.7 (c) as the simulations become inaccurate for extremely small values of $\Pr[W > w]$. Nevertheless, as in practice one only requires the knowledge of $\Pr[W > w]$ for large values of w, we feel that it is justified to state that the bounds provide a good indication of the order of magnitude of $\Pr[W > w]$ even when the approximation for $\Pr\left[\tilde{W}_2 > w\right]$ is used.

Before closing this section, we take a look at the influence of the load on the accuracy of the bounds for $\Pr[W > w]$. Therefore, these bounds are depicted versus the load in Figures 5.10-5.11. We observe that $\Pr[W > w]$ is the largest when $\rho \to 0$ and $\rho \to 1$ and that the bounds nearly coincide in these cases. Indeed, when $\rho \to 0$, few packets arrive, leading to a very long second part and a negligible short first part of the delay, whereas when $\rho \to 1$, the opposite holds. We also observe that $\Pr[W > w]$ decreases until its minimum, whereafter it increases again. In addition, the largest difference between the bounds appears in the neighbourhood of the minimum of the curves. This can be explained as follows: when ρ increases, $\Pr[W_1 > w]$ increases, whereas $\Pr[W_2 > w]$ decreases. Consequently, the difference between the bounds, $\min(\Pr[W_1 > w], \Pr[W_2 > w])$, is the largest when $\Pr[W_1 > w] = \Pr[W_2 > w]$. In that case, we learn from (5.3) and (5.4) that the upper bound is (roughly) twice as large as the lower bound. However, the order of magnitude is thus still accurately assessed, which is exactly what is required in practice.

Remark 23. *Fig. 5.9 (c) exhibits an at first sight strange jump from $w = 1$ to $w = 0$. Let us explain this. As can be observed from Fig. 5.10 (a), $\Pr[W > w]$ is dominated by $\Pr\left[\tilde{W}_2 > w\right]$ for $\rho = 0.5$ and Poisson arrivals. The jump however, stems from the approximation for $\Pr[W_1 > 0]$, which is much larger than one. This is a consequence of the approximation method based on dominant singularities, which works fine, except for very small values of w.*

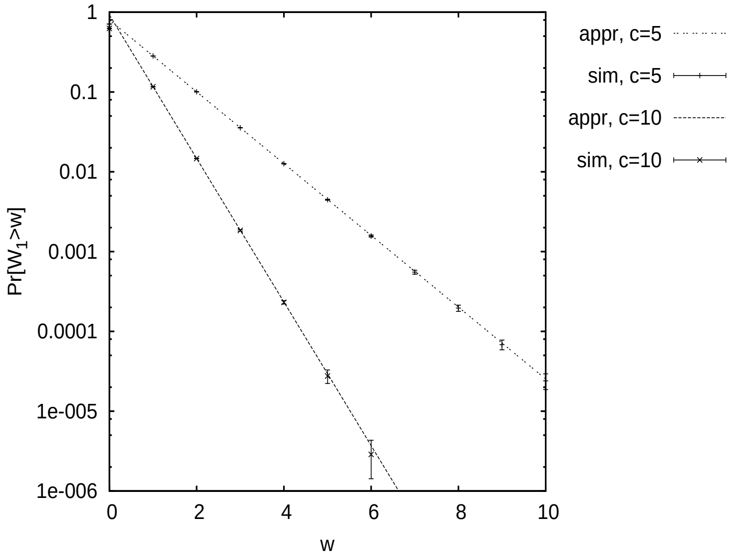


(a) several loads; $c = 5$, Poisson arrivals



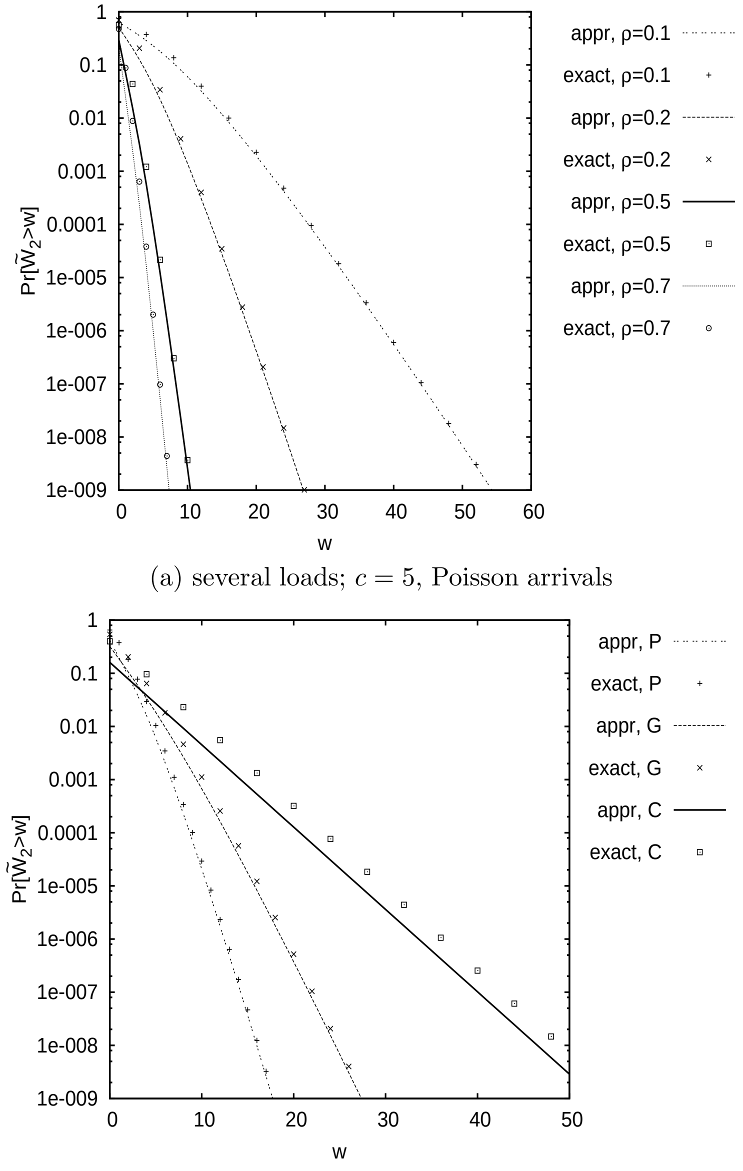
(b) several $A(z)$'s; $c = 5$, $\rho = 0.9$

Figure 5.2: Evaluation of approximation formula (5.8) for $\Pr[W_1 > w]$ (1)



(c) several c's; $\rho = 0.9$, Poisson arrivals

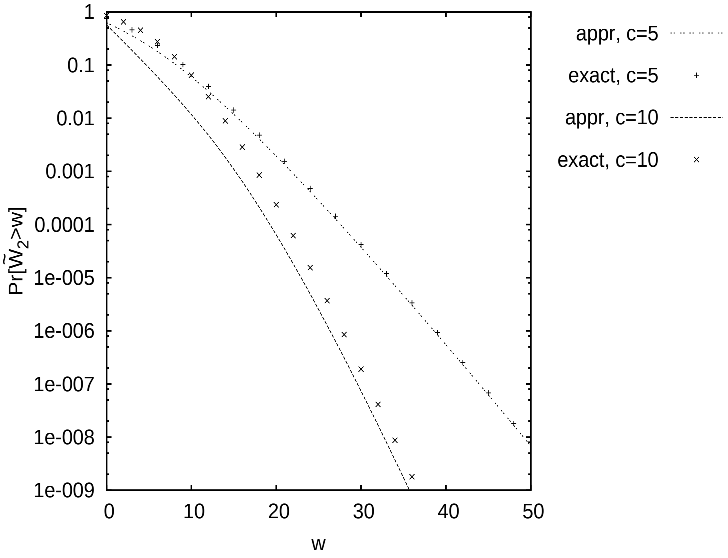
Figure 5.3: Evaluation of approximation formula (5.8) for $\Pr[W_1 > w]$ (2)



(a) several loads; $c = 5$, Poisson arrivals

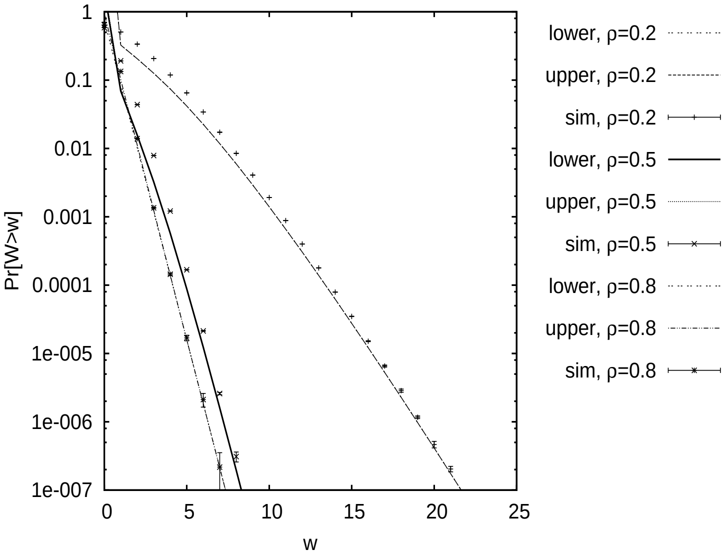
(b) several $A(z)$'s; $c = 5$, $\rho = 0.3$

Figure 5.4: Evaluation of approximation formula (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ by comparing it with exact formula (5.12) (1)

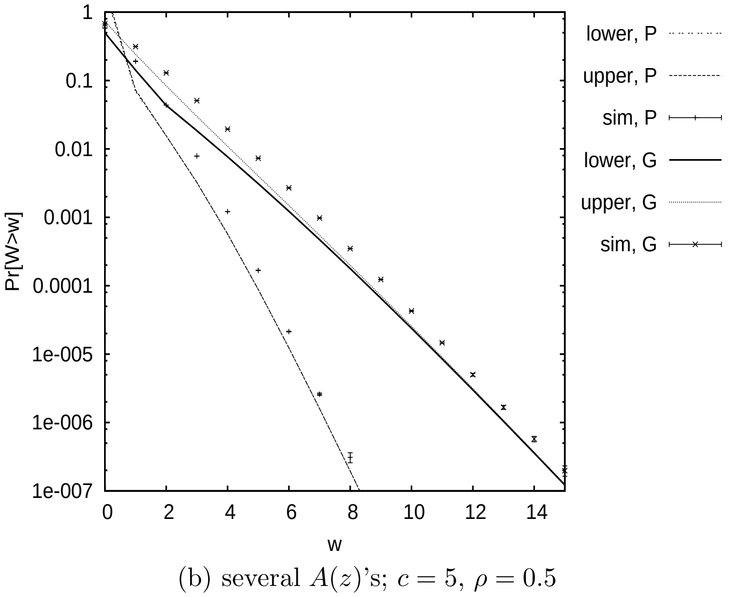


(c) several c's; $\rho = 0.3$, Poisson arrivals

Figure 5.5: Evaluation of approximation formula (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ by comparing it with exact formula (5.12) (2)

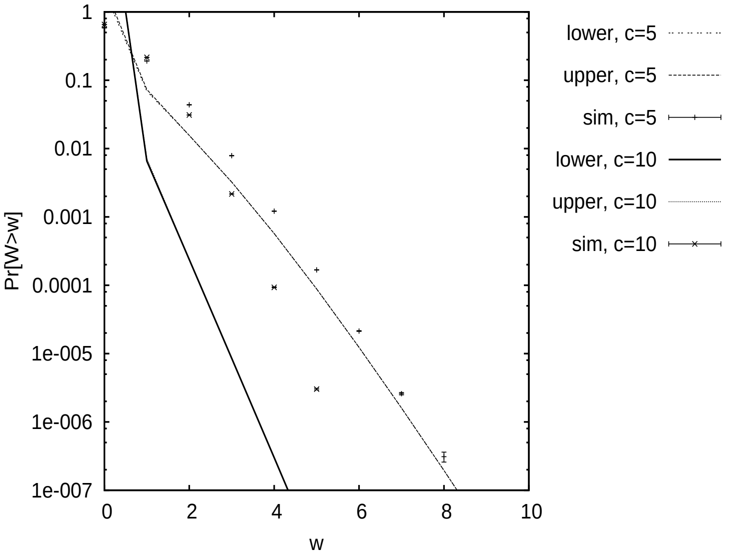


(a) several loads; $c = 5$, Poisson arrivals



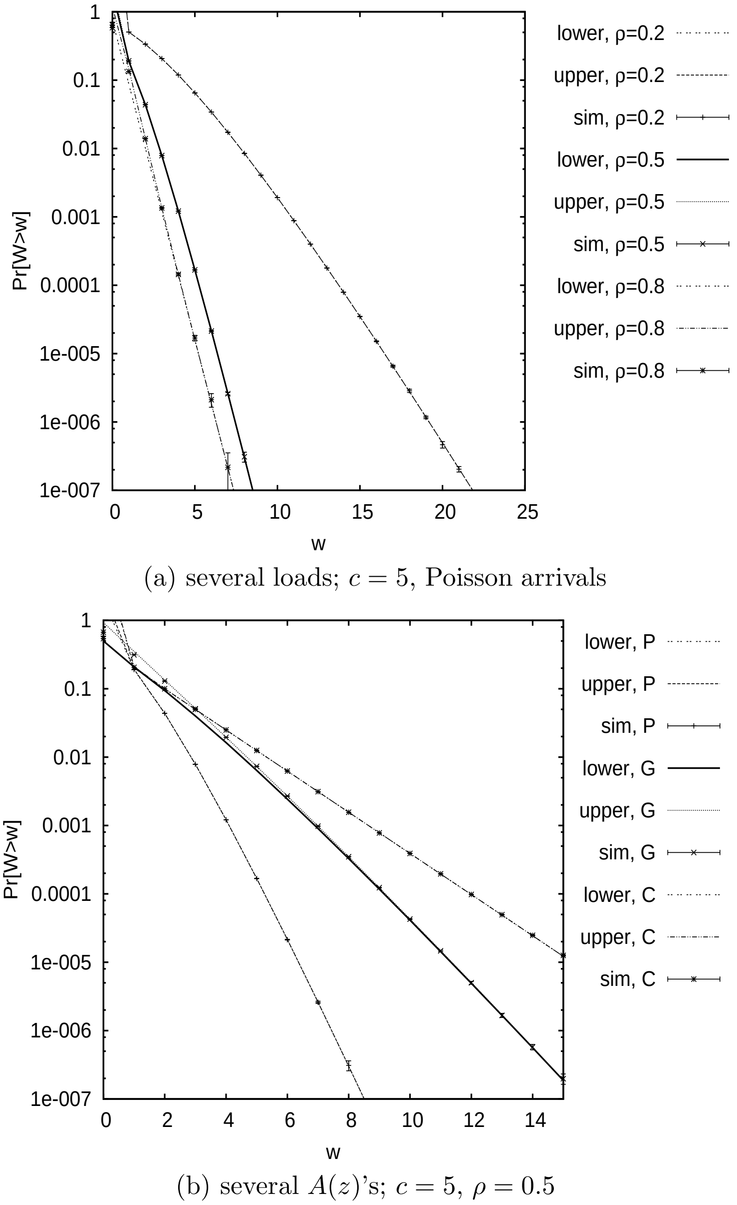
(b) several $A(z)$'s; $c = 5$, $\rho = 0.5$

Figure 5.6: Evaluation of bounds (5.3) and (5.4) for $\Pr[W > w]$; approximation formula (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (1)



(c) several c's; $\rho = 0.5$, Poisson arrivals

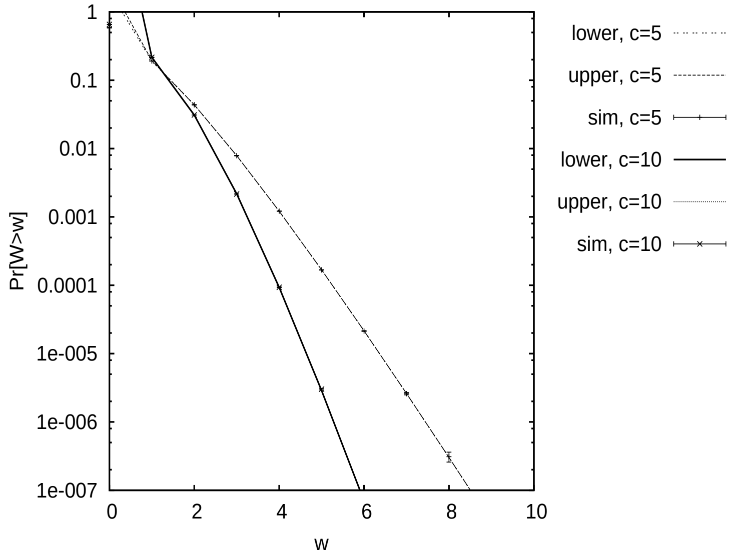
Figure 5.7: Evaluation of bounds (5.3) and (5.4) for $\Pr\left[W > w\right]$; approximation formula (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (2)



(a) several loads; $c = 5$, Poisson arrivals

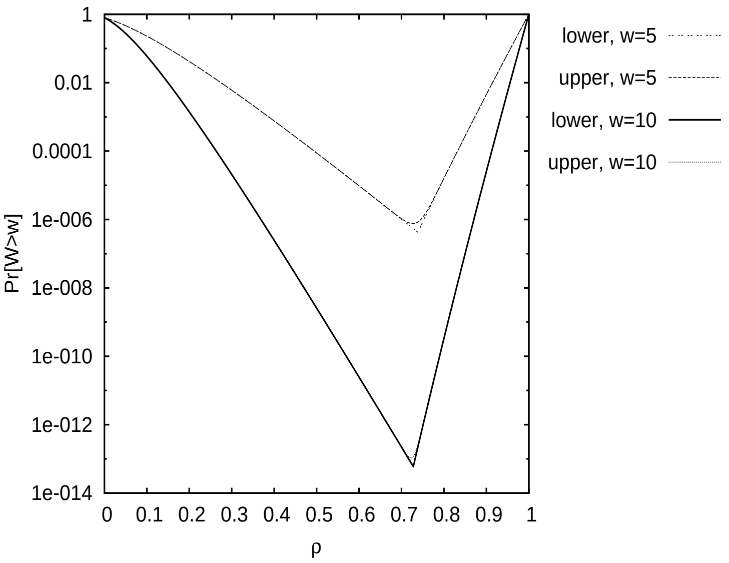
(b) several $A(z)$'s; $c = 5$, $\rho = 0.5$

Figure 5.8: Evaluation of bounds (5.3) and (5.4) for $\Pr[W > w]$; exact formula (5.12) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (1)

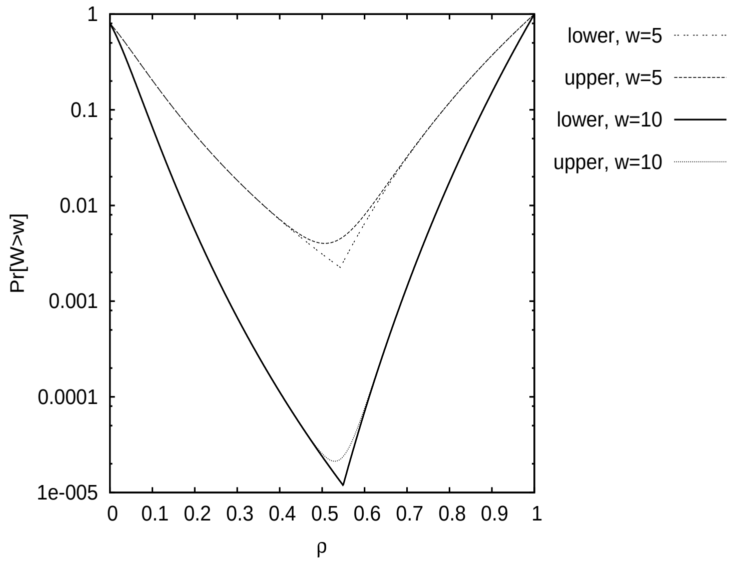


(c) several c's; $\rho = 0.5$, Poisson arrivals

Figure 5.9: Evaluation of bounds (5.3) and (5.4) for $\Pr[W > w]$; exact formula (5.12) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (2)

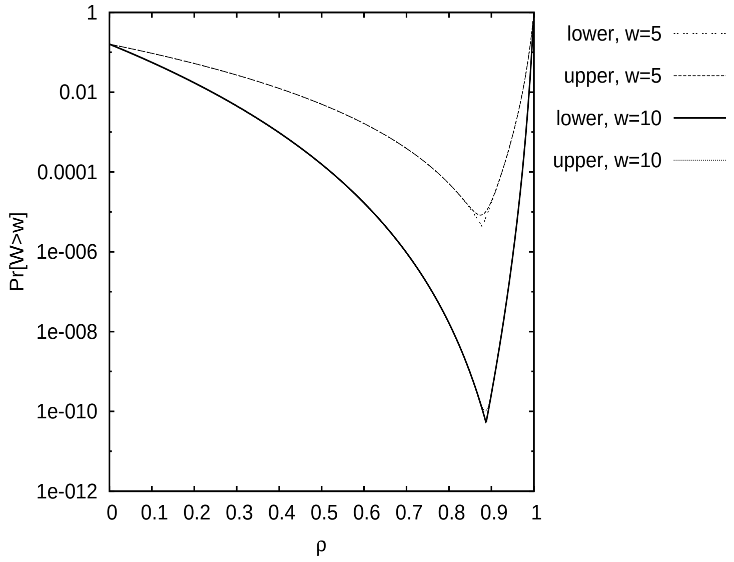


(a) Poisson



(b) geometric

Figure 5.10: Evaluation of bounds (5.3) and (5.4) for $\Pr[W > w]$; approximation formula (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ is used; $c = 5$ (1)



(c) c-centered

Figure 5.11: Evaluation of bounds (5.3) and (5.4) for $\Pr[W > w]$; approximation formula (5.18) for $\Pr\left[\tilde{W}_2 > w\right]$ is used; $c = 5$ (2)

5.3 Intermediate model

In this section, we study the tail probabilities of the delay in the intermediate model, i.e. the threshold-based model with general service times that are independent of the number of served customers ($\beta = 0$, $T_n(z) = T_c(z)\ \forall n$). Along the same lines as for the basic model, we redefine the postponing delay W_2 by shifting the starting point to the same time instant as the queueing delay W_1 and we denote it by $\tilde{W}_2$. As a result,

$$W = \max(W_1, \tilde{W}_2) \ ,$$

so that

$$\begin{aligned}\Pr\left[W > w\right] =& \Pr\left[W_1 > w \vee \tilde{W}_2 > w\right] \\ =& \Pr\left[W_1 > w\right] + \Pr\left[\tilde{W}_2 > w\right] - \Pr\left[W_1 > w \wedge \tilde{W}_2 > w\right] \ .\end{aligned}$$

Hence, we find the same bounds for $\Pr\left[W > w\right]$:

$$\Pr\left[W > w\right] \geq \max\left(\Pr\left[W_1 > w\right], \Pr\left[\tilde{W}_2 > w\right]\right) \ , \tag{5.19}$$

and

$$\Pr\left[W > w\right] \leq \Pr\left[W_1 > w\right] + \Pr\left[\tilde{W}_2 > w\right] \ . \tag{5.20}$$

These bounds again require the calculations of $\Pr\left[W_1 > w\right]$ and $\Pr\left[\tilde{W}_2 > w\right]$, which are discussed in the following subsections.

5.3.1 Calculation of $\Pr\left[W_1 > w\right]$

The PGF $W_1(z)$ is found by replacing β by 0 and $T_n(z)$ by $T_c(z)$ in (4.18):

$$\begin{aligned}W_1(z) = \frac{T_c(z) - 1}{c\lambda T_c(z)} \sum_{i=0}^{c-1} & \frac{A\left(T_c(z)^{1/c}\varepsilon_i\right) - 1}{\left(T_c(z)^{1/c}\varepsilon_i - 1\right)^2} \frac{T_c(z)^{1/c}\varepsilon_i}{z - A\left(T_c(z)^{1/c}\varepsilon_i\right)} \\ & \left\{(z-1)\sum_{n=0}^{l-1} d(n) \left(T_c(z)^{1/c}\varepsilon_i\right)^n \right. \\ & \left. + \sum_{n=l}^{c-1} d(n) \left[T_c(z) - \left(T_c(z)^{1/c}\varepsilon_i\right)^n\right]\right\} \ .\end{aligned} \tag{5.21}$$

Unlike for the basic model, considering $W_1(z^c)$ does not allow us to easily locate the dominant singularities anymore. We now search the singularity(ies) of $W_1(z)$. The singularities of $W_1(z)$ might consist of zeroes of $T_c(z)$ outside the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$, zeroes of $T_c(z)^{1/c}\varepsilon_i - 1$ outside the closed complex unit disk, zeroes of $z - A\left(T_c(z)^{1/c}\varepsilon_i\right)$ outside the complex unit disk, (possible) singularities of $T_c(z)$ and possible singularities of $A(T_c(z)^{1/c}\varepsilon_i)$. The following theorems play a crucial role in locating the dominant singularities.

Theorem 2. *Zeroes of $T_c(z)$ in the denominator of $W_1(z)$ produce no poles.*

Proof. In remark 20 (page 71), we have noticed that zeroes of $u(z,y)^c$ are no poles of $\hat{W}(z,x,y)$, whereby $u(z,y)^c \triangleq T_c(zA(y))$ and $\hat{W}(z,x,y) \triangleq \mathrm{E}\left[z^{W_1}x^G y^H\right]$, which implies that zeroes of $T_c(z)$ are no poles from $W_1(z) = \hat{W}(z,1,1)$. □

Theorem 3. *The factor $(T_c(z)^{1/c}\varepsilon_i - 1)^2$ in the denominator produces no poles for $0 \leq i \leq c-1$.*

Proof. Suppose x_i is a zero of $T_c(z)^{1/c}\varepsilon_i - 1$ with multiplicity m. Then

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k}(T_c(z)^{1/c}\varepsilon_i - 1)\bigg|_{z=x_i} = 0 \ , \qquad 0 \leq k \leq m-1 \ ,$$

and consequently:

- $\frac{\mathrm{d}^k}{\mathrm{d}z^k}(T_c(z) - 1)\Big|_{z=x_i} = 0 \ , \quad 0 \leq k \leq m-1$, meaning that x_i is also a zero of $T_c(z) - 1$ with multiplicity m, which appears in the numerator;
- $\frac{\mathrm{d}^k}{\mathrm{d}z^k}(A(T_c(z)^{1/c}\varepsilon_i) - 1)\Big|_{z=x_i} = 0 \ , \quad 0 \leq k \leq m-1$, meaning that x_i is also a zero of $A(T_c(z)^{1/c}\varepsilon_i) - 1$ with multiplicity m, which appears in the numerator.

Summarized, although x_i is a zero of $(T_c(z)^{1/c}\varepsilon_i - 1)^2$ with multiplicity $2m$, it is not a pole of $W_1(z)$, since x_i is also a zero of the numerator with multiplicity $2m$. □

Lemma 1. *Assumptions 1-5 from section 1.6 imply that $z^c - T_c(A(z))$ has exactly one zero in the interval $]1, \Re_c[$, with $\Re_c$ the radius of convergence of $T_c(A(z))$. In addition, the zero has multiplicity one and $z^c - T_c(A(z))$ contains no other zeroes with a modulus larger than one and smaller than or equal to this real zero.*

Proof. This lemma has been proved in [106]. □

Let us denote the only zero of $z^c - T_c(A(z))$ in the interval $]1, \Re_c[$ by $\tilde{z}$. Since $\tilde{z} < \Re_c \leq \Re_A$, the following definition makes sense:

$$\hat{z} \triangleq A(\tilde{z}) \ .$$

It holds that $\hat{z} \in \mathbb{R}$ and $\hat{z} > 1$, since $A(1) = 1$ and the PGF $A(z)$ is a real-valued and monotonically increasing function within $[1, \Re_A[$. In addition, $\hat{z} < \Re_{T_c}$, with $\Re_{T_c}$ the radius of convergence of $T_c(z)$, as $\tilde{z} < \Re_c$ implies that $\hat{z} = A(\tilde{z}) < \Re_{T_c}$.

Theorem 4. *Assumptions 1-5 imply that*

1. *$T_c(\hat{z})^{1/c} < \Re_A$ and $\hat{z}$ is a zero of $z - A(T_c(z)^{1/c})$;*
2. *the equations $z - A(T_c(z)^{1/c}\varepsilon_i) \ , \ 0 \leq i \leq c-1$ contain no other zeroes with a modulus larger than one and smaller than or equal to $\hat{z}$;*
3. *$\hat{z}$ is a zero of $z - A(T_c(z)^{1/c})$ with multiplicity one.*

Proof. 1. On account of lemma 1, we have

$$\tilde{z}^c = T_c(A(\tilde{z})) \ . \tag{5.22}$$

As $\tilde{z}$ and $T_c(A(\tilde{z}))$ are both real positive numbers, (5.22) can be transformed into

$$\tilde{z} = T_c(A(\tilde{z}))^{1/c} \ ,$$

which is, owing to the definition of $\hat{z}$, equivalent to

$$\tilde{z} = T_c(\hat{z})^{1/c} \ .$$

Finally, taking into account that $T_c(\hat{z})^{1/c} = \tilde{z} < \Re_c \leq \Re_A$ and invoking the definition of $\hat{z}$, we find

$$\hat{z} = A(T_c(\hat{z})^{1/c}) \ .$$

In other words, $T_c(\hat{z})^{1/c} < \Re_A$ and $\hat{z}$ is a zero of $z - A(T_c(z)^{1/c})$.

2. This part is a proof by contradiction. Assume that an i ($0 \leq i \leq c-1$) exists, for which $z - A(T_c(z)^{1/c}\varepsilon_i)$ has a zero, z^*, with $z^* \neq \hat{z}$ and $1 < |z^*| \leq \hat{z}$. Owing to $|z^*| \leq \hat{z} < \Re_{T_c}$, the following definition makes sense: $\tilde{z}^* \triangleq T_c(z^*)^{1/c}\varepsilon_i$. Consequently, we have that

$$|\tilde{z}^*|^c = |T_c(z^*)| \leq \sum_{n=1}^{\infty} \Pr[T_c = n]\, |z^*|^n \leq \sum_{n=1}^{\infty} \Pr[T_c = n]\, \hat{z}^n = T_c(\hat{z}) = \tilde{z}^c \ .$$

Hence, as both $|\tilde{z}^*|$ and $\tilde{z}$ are positive real numbers,

$$|\tilde{z}^*| \leq \tilde{z} \ . \tag{5.23}$$

This implies that $|\tilde{z}^*| < \Re_A$ and taking into account that $z^* = A(T_c(z^*)^{1/c}\varepsilon_i)$, we find that $z^* = A(\tilde{z}^*)$. As a consequence, $|A(\tilde{z}^*)| < \Re_{T_c}$ and $T_c(A(\tilde{z}^*)) = T_c(z^*) = (\tilde{z}^*)^c$, meaning that $\tilde{z}^*$ is a zero of $z^c - T_c(A(z))$. On account of lemma 1 however, we have that $\tilde{z}$ is the zero with the smallest modulus larger than one of this equation and $\tilde{z}$ is the only zero with that modulus, so that $|\tilde{z}^*| > \tilde{z}$, which is a contradiction with (5.23).

3. The property of $\hat{z}$ having multiplicity one is also a proof by contradiction. If $\hat{z}$ would have a multiplicity larger than one, then (we use primes to indicate derivatives)

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}z}[z - A(T_c(z)^{1/c})]\bigg|_{z=\hat{z}} = 0 \\ \Leftrightarrow\ & 1 - A^{'}(T_c(\hat{z})^{1/c})\frac{1}{c}T_c(\hat{z})^{1/c-1}T_c^{'}(\hat{z}) = 0 \ . \end{aligned}$$

Writing this in terms of $\tilde{z}$ instead of in $\hat{z}$ and relying on $\tilde{z}^c = T_c(A(\tilde{z}))$, we further transform this to

$$\begin{aligned} & c - A^{'}(\tilde{z})\tilde{z}^{1-c}T_c^{'}(A(\tilde{z})) = 0 \\ \Leftrightarrow\ & c\tilde{z}^{c-1} - T_c^{'}(A(\tilde{z}))A^{'}(\tilde{z}) = 0 \\ \Leftrightarrow\ & \frac{\mathrm{d}}{\mathrm{d}z}[z^c - T_c(A(z))]\bigg|_{z=\tilde{z}} = 0 \ , \end{aligned}$$

meaning that $\tilde{z}$ is a zero of $z^c - T_c(A(z))$ with multiplicity larger than one, which is impossible according to lemma 1. □

Summarized, $W_1(z)$ has one dominant singularity, being the pole $\hat{z}$. This dominant pole is a zero of $z - A(T_c(z)^{1/c})$, it has multiplicity one and is equal to $A(\tilde{z})$, with $\tilde{z}$ the only zero in $]1, \Re_c[$ of $z^c - T_c(A(z))$. As $\tilde{z} \in \mathbb{R}$, it can be easily determined numerically, for instance with the bisection or the Newton-Raphson method.
Taking these findings into account, we find that $W_1(z)$ is in the neighborhood of $\hat{z}$ proportional to

$$\begin{aligned} W_1(z) \sim & \frac{T_c(z) - 1}{c\lambda T_c(z)} \frac{\left[A\left(T_c(z)^{1/c}\right) - 1\right] T_c(z)^{1/c}}{\left[z - A\left(T_c(z)^{1/c}\right)\right] \left[T_c(z)^{1/c} - 1\right]^2} \\ & \left\{(z-1)\sum_{n=0}^{l-1} d(n) T_c(z)^{n/c} + \sum_{n=l}^{c-1} d(n) \left[T_c(z) - T_c(z)^{n/c}\right]\right\} \ . \end{aligned}$$

Consequently, application of formula (1.4) of Darboux's theorem yields

$$\begin{aligned} & \Pr\left[W_1 > w\right] \\ \approx \quad & \frac{\hat{z}^{-(w+1)}}{1-\hat{z}} \frac{T_c(\hat{z}) - 1}{\lambda} \frac{A\left(T_c(\hat{z})^{1/c}\right) - 1}{\left(T_c(\hat{z})^{1/c} - 1\right)^2} T_c(\hat{z})^{\frac{1}{c}-1} \\ & \cdot \frac{(\hat{z}-1)\sum_{n=0}^{l-1} d(n) T_c(\hat{z})^{n/c} + \sum_{n=l}^{c-1} d(n) \left(T_c(\hat{z}) - T_c(\hat{z})^{n/c}\right)}{c - A'\left(T_c(\hat{z})^{1/c}\right) T_c(\hat{z})^{\frac{1}{c}-1} T_c'(\hat{z})} \ . \end{aligned} \tag{5.24}$$

5.3.2 Calculation of $\Pr\left[\tilde{W}_2 > w\right]$

In order to calculate $\Pr\left[\tilde{W}_2 > w\right]$, we start from the following relation:

$$\Pr\left[\tilde{W}_2 > w\right] = \Pr\left[\left(\left[Q_J + A^-\right] \bmod c\right) + 1 + A^+ + \sum_{i=1}^{w} A_{J+i} < l\right] \ . \tag{5.25}$$

This relation is the same as for the basic model, except that the number of customers has to reach the service threshold l instead of the server capacity c. Analogously as in the previous section, we find that

$$\mathrm{E}\left[x^{\left([Q_J + A^-] \bmod c\right)} x^{A^+}\right] = \frac{x^c - 1}{c(x-1)} \sum_{i=0}^{c-1} Q(\varepsilon_i) \frac{A(\varepsilon_i) - A(x)}{\lambda(\varepsilon_i - x)} \frac{\varepsilon_i(x-1)}{x - \varepsilon_i} \ , \tag{5.26}$$

with $Q(z)$ the PGF of the queue content at a random slot mark, which is found by setting $\beta = 0$ and $T_n(z) = T_c(z)$ in (2.17):

$$Q(z) = \frac{1}{z^c - T_c(A(z))} \left\{(z^c - 1)\sum_{n=0}^{l-1} d(n) z^n + \frac{T_c(A(z)) - 1}{A(z) - 1} \sum_{n=l}^{c-1} d(n)(z^c - z^n)\right\} \ ,$$

implying that $Q(\varepsilon_0) = 1$ (since $\varepsilon_0 = 1$) and

$$Q(\varepsilon_i) = \frac{\sum_{n=l}^{c-1} d(n)(\varepsilon_i^n - 1)}{A(\varepsilon_i) - 1} \ , \qquad 1 \leq i \leq c-1 \ .$$

Relying on this result, we find:

$$\mathrm{E}\left[x^{\left([Q_J + A^-] \bmod c\right)} x^{A^+}\right] = \frac{x^c - 1}{c(x-1)} f(x) \ , \tag{5.27}$$

with

$$f(x) = \frac{1-A(x)}{\lambda(1-x)} + \sum_{i=1}^{c-1} \frac{A(\varepsilon_i)-A(x)}{\lambda(\varepsilon_i - x)} \frac{\varepsilon_i(x-1)}{x-\varepsilon_i} \frac{\sum_{n=l}^{c-1} d(n)(\varepsilon_i^n - 1)}{A(\varepsilon_i)-1} \ ,$$

with the first term in the right-hand-side the one for $i = 0$. The combination of (5.25), (5.27) and the probability generating property of PGFs produces

$$\begin{aligned} \Pr\left[\tilde{W}_2 > w\right] &= \sum_{m=0}^{l-1} \frac{1}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} \mathrm{E}\left[x^{([Q_J+A^-] \bmod c)+1+A^+ + \sum_{i=1}^{w} A_{J+i}}\right]\Big|_{x=0} \\ &= \sum_{m=1}^{l-1} \frac{1}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} x A(x)^w \frac{x^c-1}{c(x-1)} f(x)\Big|_{x=0} \ . \end{aligned}$$

After some mathematical manipulations, this can be transformed into

$$\Pr\left[\tilde{W}_2 > w\right] = \sum_{m=0}^{l-2} \frac{1}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} A(x)^w \frac{x^c-1}{c(x-1)} f(x)\Big|_{x=0}$$

$$= \frac{1}{c} \sum_{m=0}^{l-2} \sum_{k=0}^{m} \frac{1}{k!(m-k)!} \frac{\mathrm{d}^k}{\mathrm{d}x^k} \frac{x^c-1}{x-1}\Big|_{x=0} \frac{\mathrm{d}^{m-k}}{\mathrm{d}x^{m-k}} A(x)^w f(x)\Big|_{x=0} \ .$$

As

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k} \frac{x^c-1}{x-1}\Big|_{x=0} = k! \ , \qquad k < c \ ,$$

we find

$$\begin{aligned} \Pr\left[\tilde{W}_2 > w\right] &= \frac{1}{c} \sum_{m=0}^{l-2} \sum_{k=0}^{m} \frac{1}{(m-k)!} \frac{\mathrm{d}^{m-k}}{\mathrm{d}x^{m-k}} A(x)^w f(x)\Big|_{x=0} \\ &= \frac{1}{c} \sum_{m=0}^{l-2} \frac{l-1-m}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} A(x)^w f(x)\Big|_{x=0} \ . \end{aligned} \tag{5.28}$$

Formula (5.28) can be implemented in a mathematical program such as matlab. This procedure suffers from the drawback that high-order derivatives may have to be computed, which causes a considerable reduction in speed and even is infeasible if l and c are quite large. Therefore, we now deduce an approximation for $\Pr\left[\tilde{W}_2 > w\right]$, whereby no derivatives have to be taken.

Multiplying both sides of (5.28) by z^w and taking the sum over all values of w produces

$$\sum_{w=0}^{\infty} \Pr\left[\tilde{W}_2 > w\right] z^w = \frac{\tilde{W}_2(z)-1}{z-1} = \frac{1}{c} \sum_{m=0}^{l-2} \frac{l-1-m}{m!} \frac{\partial^m}{\partial x^m} \frac{f(x)}{1-zA(x)}\Big|_{x=0} \ . \tag{5.29}$$

The m-th ($m \geq 0$) derivative of $f(x)/(1 - zA(x))$ can be written as

$$\frac{\partial^m}{\partial x^m} \frac{f(x)}{1-zA(x)} = \sum_{j=0}^{m} \frac{C_{m,j}(z,x)}{[1-zA(x)]^{j+1}} \ , \tag{5.30}$$

whereby $C_{m,j}(z, x)$ are functions of z and x that have no factor $1 - zA(x)$ in the denominator. As opposed to $C_{m,j}(z, x)$ for $j \neq m$, $C_{m,m}(z, x)$ is relatively easy to calculate:

$$C_{m,m}(z,x) = m! f(x) z^m A^{'}(x)^m \ .$$

The substitution of (5.30) in (5.29) yields

$$\frac{\tilde{W}_2(z)-1}{z-1}=\frac{1}{c}\sum_{m=0}^{l-2}\frac{l-1-m}{m!}\sum_{j=0}^{m}\frac{C_{m,j}(z,0)}{[1-zA(0)]^{j+1}}\ . \tag{5.31}$$

From this equation, it is clear that $z=1/A(0)$ is the dominant pole of $[\tilde{W}_2(z)-1]/(z-1)$ and that it has multiplicity $l-1$. Analogously as in the previous section, we retain for every m the term that produces the largest power of $1-zA(0)$ in the denominator. We thus take advantage of the fact that we can easily calculate $C_{m,m}(z,x)$ for all m. Hence, in the neighborhood of $z=1/A(0)$, $[\tilde{W}_2(z)-1]/(z-1)$ is proportional to

$$\frac{\tilde{W}_2(z)-1}{z-1}\sim\frac{1}{c}\sum_{m=0}^{l-2}\frac{(l-1-m)f(0)z^m A^{'}(0)^m}{[1-zA(0)]^{m+1}}\ . \tag{5.32}$$

Finally, recall that we have deduced in section 5.2 that

$$\frac{1}{[1-zA(0)]^{m+1}}=\frac{1}{m!A(0)^m}\sum_{w=m}^{\infty}A(0)^w z^{w-m}\frac{w!}{(w-m)!}\ . \tag{5.33}$$

As a result, the substitution of (5.33) in (5.32) produces:

$$\begin{aligned}\frac{\tilde{W}_2(z)-1}{z-1}&\sim\frac{f(0)}{c}\sum_{m=0}^{l-2}A^{'}(0)^m(l-1-m)\sum_{w=m}^{\infty}z^w\frac{w!}{m!(w-m)!}A(0)^{w-m}\\&=\frac{f(0)}{c}\sum_{w=0}^{\infty}z^w\sum_{m=0}^{\min(l-2,w)}A^{'}(0)^m(l-1-m)\binom{w}{m}A(0)^{w-m}\ .\end{aligned}$$

Equating powers of z^w at both sides of the equation and taking into account that $[\tilde{W}_2(z)-1]/(z-1)=\sum_{w=0}^{\infty}\Pr\left[\tilde{W}_2>w\right]z^w$ finally yields

$$\Pr\left[\tilde{W}_2>w\right]\approx\frac{f(0)}{c}\sum_{m=0}^{\min(l-2,w)}A^{'}(0)^m(l-1-m)\binom{w}{m}A(0)^{w-m}\ . \tag{5.34}$$

Note that for large w, formula (5.34) becomes a sum from 0 to $l-2$. We further point out that the binomial coefficient causes no difficulties, since efficient routines exist to calculate them, even for large w.

Remark 24. *Note again that this approach is not suited for cases whereby $A^{'}(0)=0$, as only the term corresponding to $m=0$ in (5.32) differs from 0. In these cases, additional terms with $j<m$ must be taken into account from (5.31).*

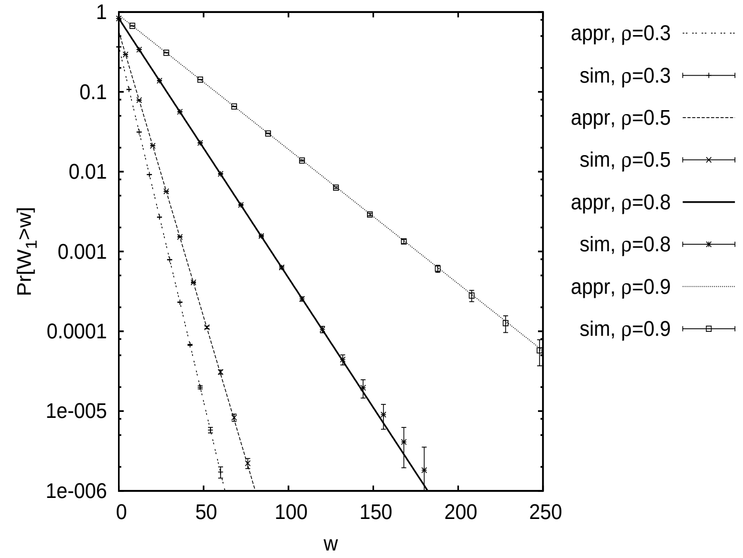
Remark 25. *When $l=c$, $f(0)=(1-A(0))/\lambda$ and (5.34) reduces to formula (5.18) for $\Pr\left[\tilde{W}_2>w\right]$ in case of the basic model. Note also that when $l=c$, $\Pr\left[\tilde{W}_2>w\right]$ is independent of $T_c(z)$. The reason is that the postponing delay of a customer depends on the position of the customer in its served batch (as it determines how many customers still have to arrive) and the future arrivals. When $l=c$, the server always serves full batches so that the position is uniformly distributed between 1 and c and the future arrivals are of course also independent of the service times.*

5.3.3 Evaluation of approximation formulas

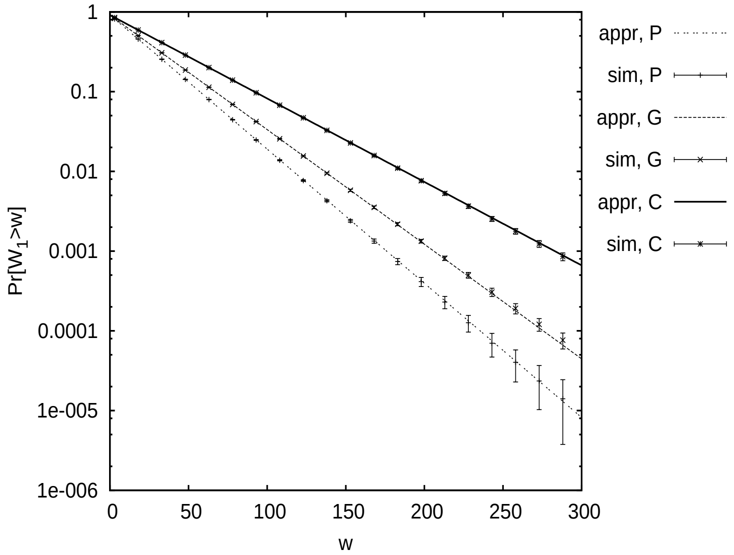
In this section, we evaluate the accuracy of our approach. First, we study formula (5.24) for $\Pr[W_1 > w]$. Then, we focus on approximation (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ and finally the accuracy of the bounds for $\Pr[W > w]$ is covered.

In Figures 5.12-5.14, we have depicted approximation (5.24) as well as simulated values[2] for $\Pr[W_1 > w]$ versus w for various combinations of distributions for the number of customer arrivals (Poisson $A(z) = e^{\lambda(z-1)}$; Geometric $A(z) = 1/(1 + \lambda - \lambda z)$; C-center $A(z) = 1 - \lambda/c + \lambda/(2c)(z^{c-1} + z^{c+1})$) and service times (Geometric $T_c(z) = z/[\mathrm{E}[T_c] + (1 - \mathrm{E}[T_c])z]$; 25 $T_c(z) = (25 - \mathrm{E}[T_c])z/24 + (\mathrm{E}[T_c] - 1)z^{25}/24$ with $\mathrm{E}[T_c] = 5$ or 10) and several server capacities, service thresholds and loads. We can draw the same conclusions as for the basic model: the approximation is accurate, even for smaller values of w, a higher load leads to larger tail probabilities whereas a larger server capacity has a beneficial effect even when the load remains equal (and thus the mean arrival rate λ is larger). In addition, we perceive that a larger variance in the arrival and service process causes slower decaying probabilities and from Fig. 5.14 we deduce that the service threshold has nearly no impact.
Next, approximation (5.34) and exact formula (5.28) for $\Pr\left[\tilde{W}_2 > w\right]$ are depicted versus w in Figures 5.15-5.17 for various settings of the system parameters. Along the same lines as for the basic model, we observe that the approximation is accurate for larger values of w and thus that it is extremely suited for quickly assessing the order of magnitude of $\Pr\left[\tilde{W}_2 > w\right]$. Furthermore, an increasing load as well as an increasing server capacity c (even when ρ remains equal) has a positive influence on the probabilities. Figure 5.17 also exhibits that $\Pr\left[\tilde{W}_2 > w\right]$ decays slower in case of a larger service threshold l. Finally, when $l = 1$, $\Pr\left[\tilde{W}_2 > w\right]$ equals zero as the server then immediately starts a new service when being available and finding a customer.
As a closer, we evaluate the lower and upper bounds (5.19) and (5.20) in Figures 5.18-5.22. We again observe that the approximations are accurate for large w. Furthermore, we perceive that a small range in the load exists where the bounds differ somewhat. This is, as explained for the basic model, the area where $\Pr[W_1 > w] \approx \Pr\left[\tilde{W}_2 > w\right]$. Examination of the bounds shows that in that case the upper bound is more or less twice the lower bound. Anyway, based on these results we feel that it is justified to conclude that the approximations are extremely suited for assessing the order of magnitude of $\Pr[W > w]$ for large values of w.

[2] Only for $\Pr\left[\tilde{W}_2 > w\right]$ we have an exact formula at our disposal. For the other tail probabilities ($\Pr[W_1 > w]$ and $\Pr[W > w]$) throughout this section, we have therefore depicted the 95% confidence intervals resulting from 10 Monte Carlo simulations whereby each simulation generates W_1 and W for 10^8 customers.

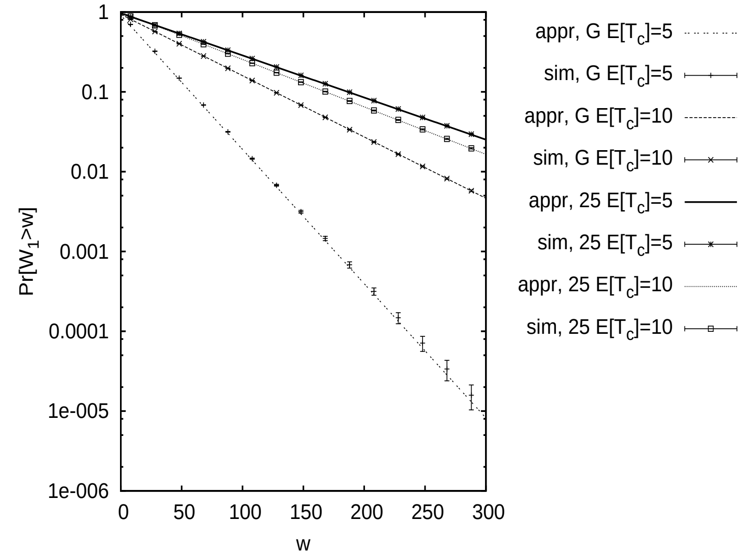


(a) several loads; $c = 5$, $l = 3$, Poisson arrivals, geometric services $\mathrm{E}[T_c] = 5$

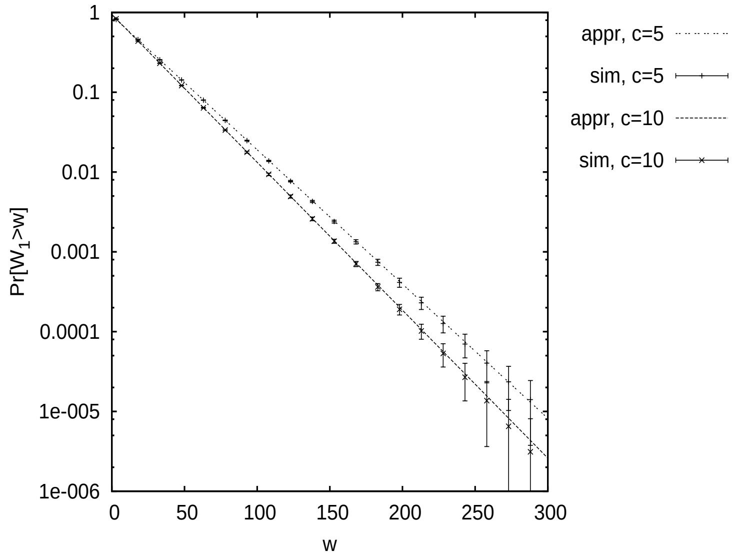


(b) several $A(z)$'s; $c = 5$, $l = 3$, $\rho = 0.9$, geometric services $\mathrm{E}[T_c] = 5$

Figure 5.12: Evaluation of approximation formula (5.24) for $\Pr[W_1 > w]$ (1)

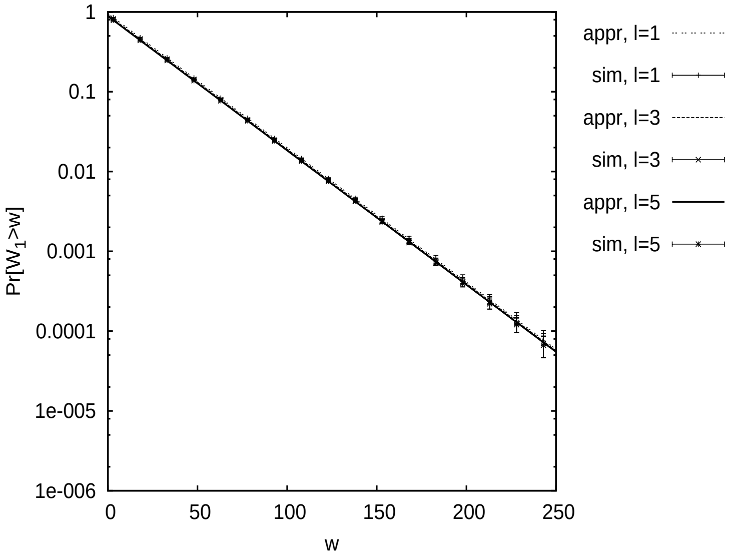


(c) several $T_c(z)$'s; $c = 5$, $l = 3$, $\rho = 0.9$, Poisson arrivals



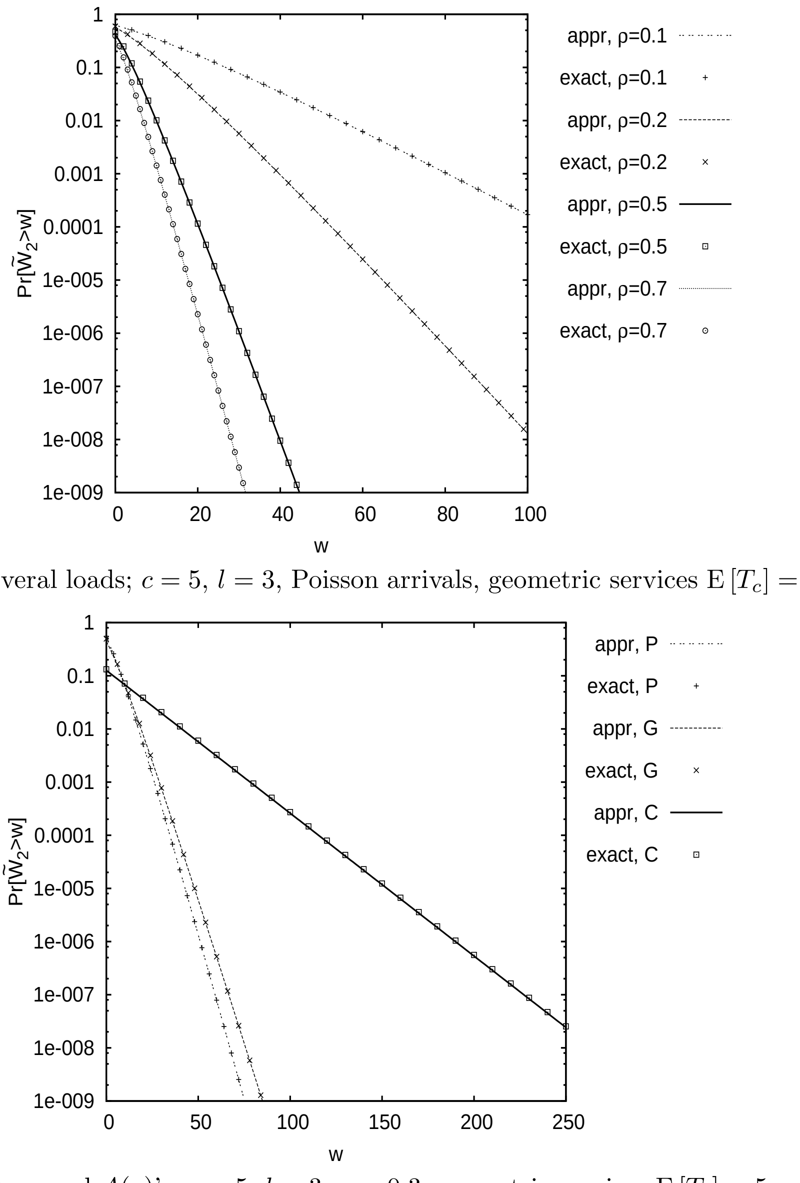
(d) several c's; $l = 3$, $\rho = 0.9$, Poisson arrivals, geometric services $\mathrm{E}[T_c] = 5$

Figure 5.13: Evaluation of approximation formula (5.24) for $\Pr[W_1 > w]$ (2)



(e) several l's; $c = 5$, $\rho = 0.9$, Poisson arrivals, geometric services $\mathrm{E}\left[T_c\right] = 5$

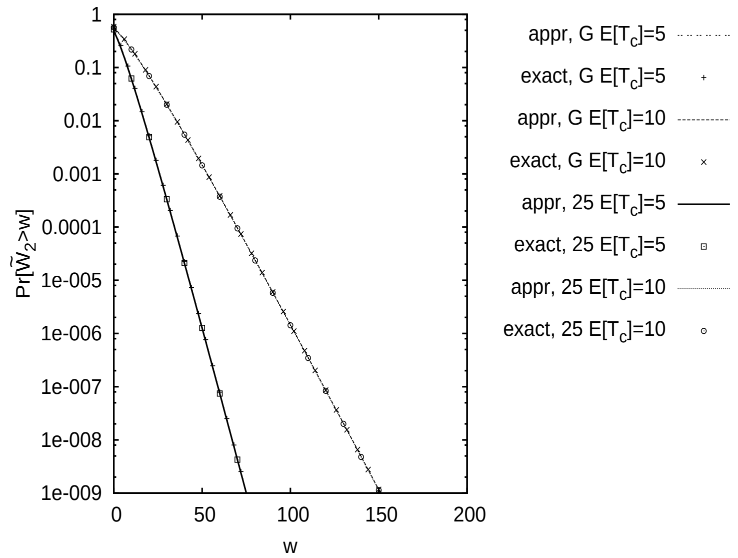
Figure 5.14: Evaluation of approximation formula (5.24) for $\Pr\left[W_1 > w\right]$ (3)



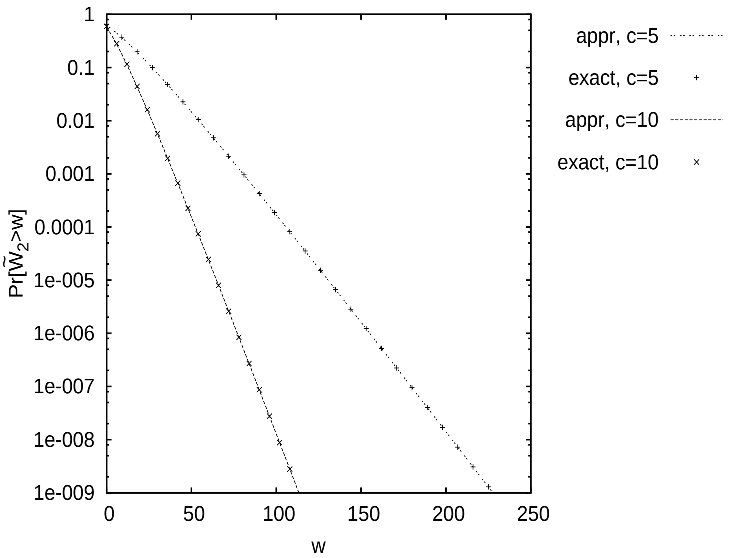
(a) several loads; $c = 5$, $l = 3$, Poisson arrivals, geometric services $\mathrm{E}[T_c] = 5$

(b) several $A(z)$'s; $c = 5$, $l = 3$, $\rho = 0.3$, geometric services $\mathrm{E}[T_c] = 5$

Figure 5.15: Evaluation of approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ by comparing it with exact formula (5.28) (1)

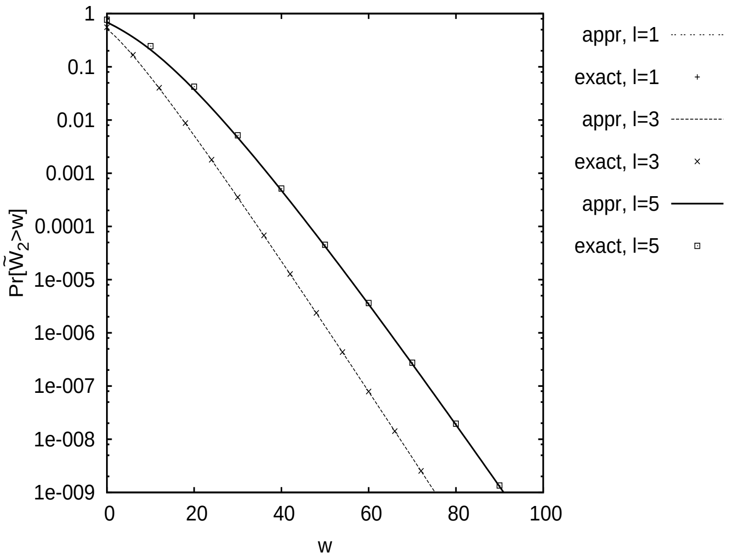


(c) several $T_c(z)$'s; $c = 5$, $l = 3$, $\rho = 0.3$, Poisson arrivals



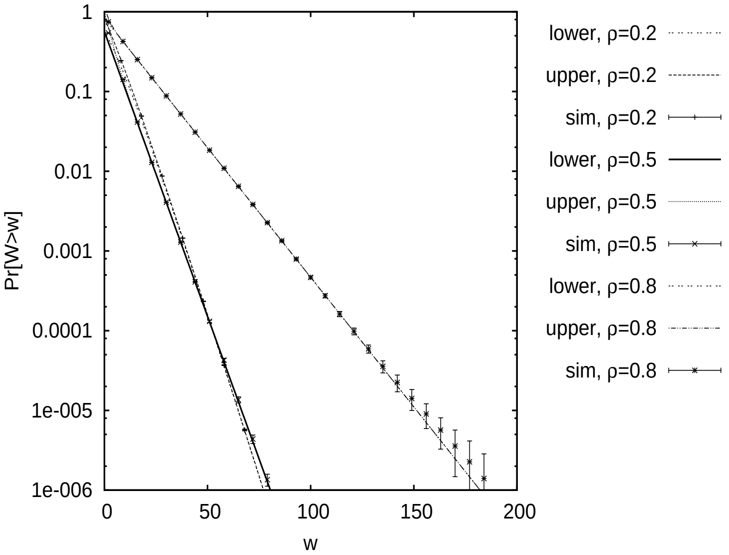
(d) several c's; $l = 3$, $\rho = 0.3$, Poisson arrivals, geometric services $\mathrm{E}[T_c] = 5$

Figure 5.16: Evaluation of approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ by comparing it with exact formula (5.28) (2)

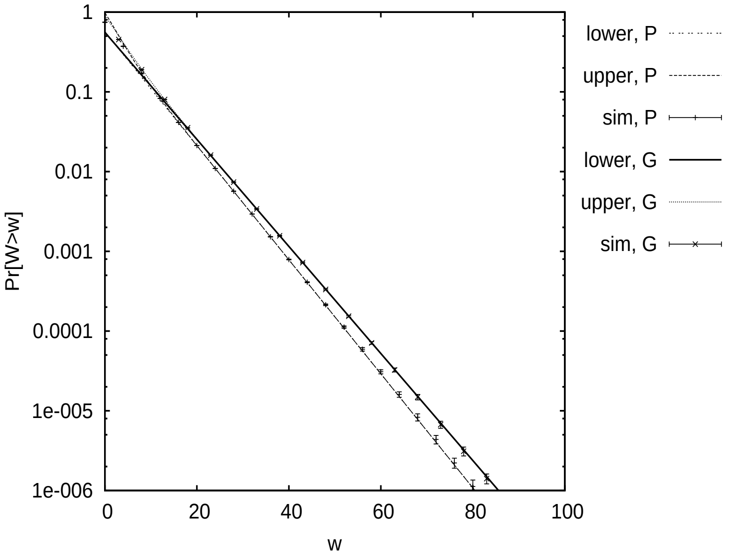


(e) several l's; $c = 5$, $\rho = 0.3$, Poisson arrivals, geometric services $\mathrm{E}[T_c] = 5$

Figure 5.17: Evaluation of approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ by comparing it with exact formula (5.28) (3)

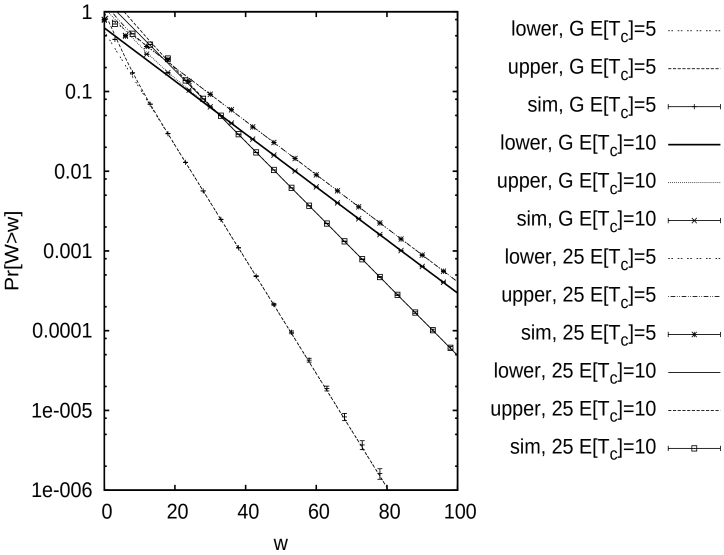


(a) several loads; $c = 5$, $l = 3$, Poisson arrivals, geometric services $\mathrm{E}\left[T_c\right] = 5$

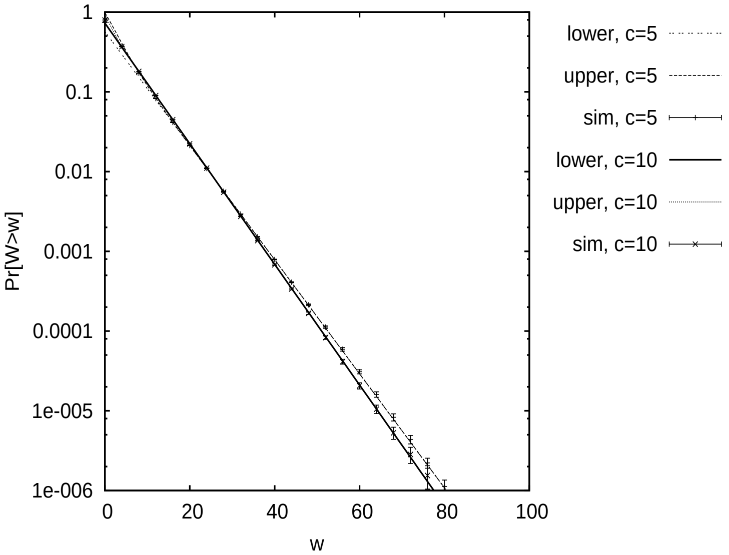


(b) several $A(z)$'s; $c = 5$, $l = 3$, $\rho = 0.5$, geometric services $\mathrm{E}\left[T_c\right] = 5$

Figure 5.18: Evaluation of bounds for $\Pr\left[W > w\right]$; approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (1)

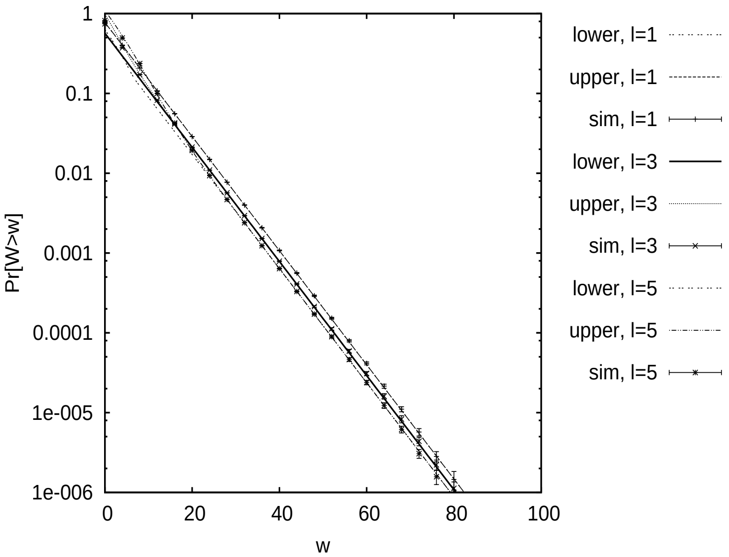


(c) several $T_c(z)$'s; $c = 5$, $l = 3$, $\rho = 0.5$, Poisson arrivals



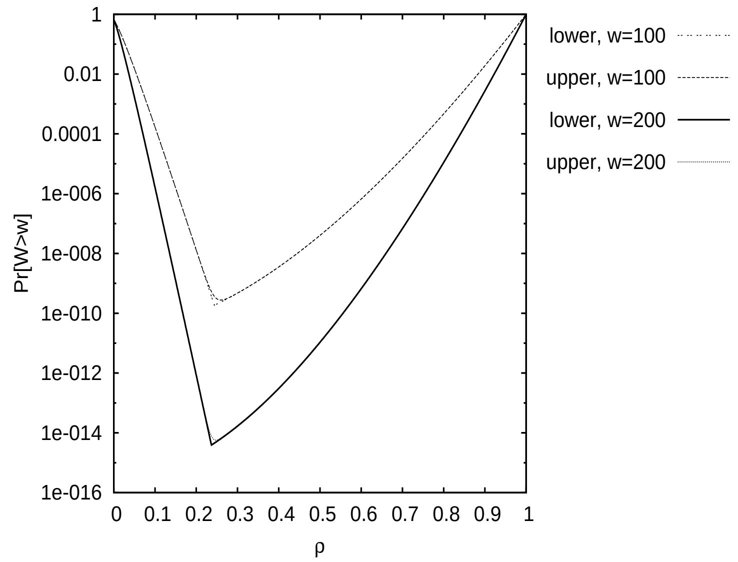
(d) several c's; $l = 3$, $\rho = 0.5$, Poisson arrivals, geometric services $\mathrm{E}[T_c] = 5$

Figure 5.19: Evaluation of bounds for $\Pr[W > w]$; approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (2)

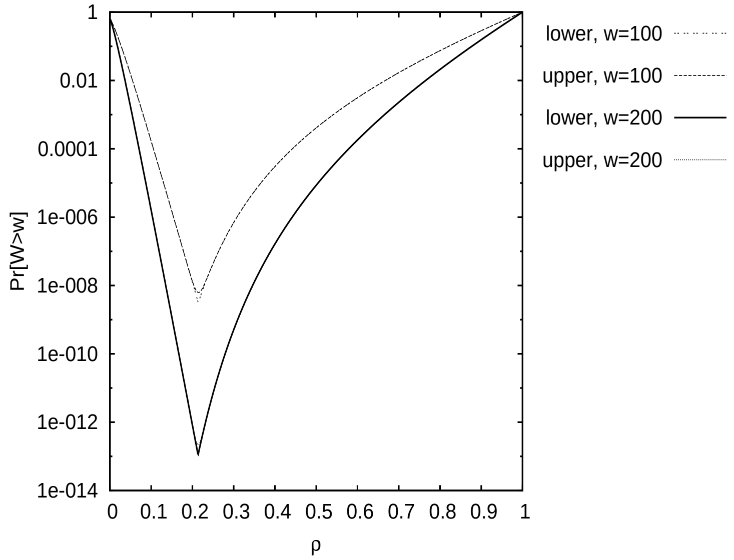


(e) several l's; $c = 5$, $\rho = 0.5$, Poisson arrivals, geometric services $\mathrm{E}\left[T_c\right] = 5$

Figure 5.20: Evaluation of bounds for $\Pr\left[W > w\right]$; approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ is used (3)

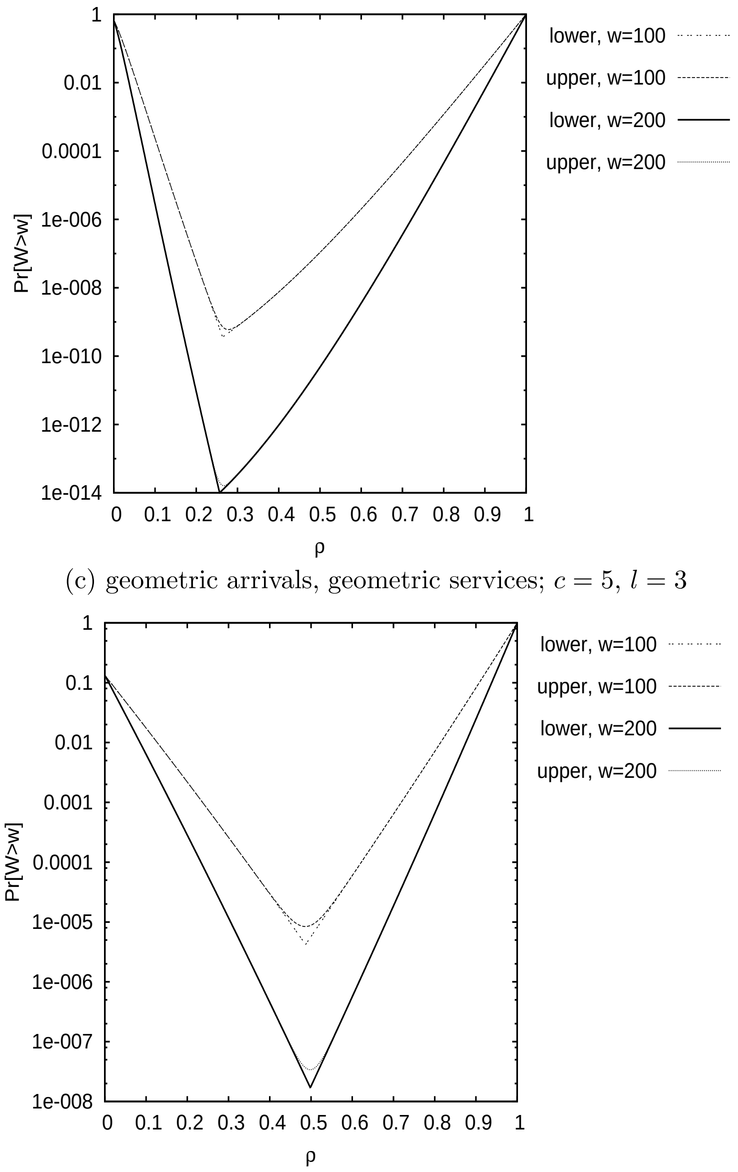


(a) Poisson arrivals, geometric services; $c = 5$, $l = 3$



(b) Poisson arrivals, 1 or 25 slots service; $c = 5$, $l = 3$

Figure 5.21: Evaluation of bounds for $\Pr[W > w]$; approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ is used; $c = 5$, $\mathrm{E}[T_c] = 5$ (1)



(c) geometric arrivals, geometric services; $c = 5$, $l = 3$

(d) c-centered arrivals, geometric services; $c = 5$, $l = 3$

Figure 5.22: Evaluation of bounds for $\Pr[W > w]$; approximation formula (5.34) for $\Pr\left[\tilde{W}_2 > w\right]$ is used; $c = 5$, $\mathrm{E}[T_c] = 5$ (2)

5.4 Final model

In this section, we study the tail probabilities of the customer delay for the final model, whereby service times are dependent on the number of served customers and with the timer mechanism controlled by β. When the server becomes available and finds less than l customers, a new service might nevertheless be started with probability β. At first sight, it seems evident to deduce approximations for $\Pr[W > w]$ completely analogously as for the previous models. Doing so, we would find that

$$W = \max\left(W_1, \tilde{W}_2\right) \quad ,$$

whereby $\tilde{W}_2$ represents the redefined postponing delay, so that the same approximations would be obtained. In order to compute $\Pr\left[\tilde{W}_2 > w\right]$, we could start from the following relation:

$$\Pr\left[\tilde{W}_2 > w\right] = \Pr\left[\left([Q_J + A^-] \bmod c\right) + 1 + A^+ + \sum_{i=1}^{w} A_{J+i} < l\right](1-\beta)^{w+1} \quad , \qquad (5.35)$$

with J the tagged customer's arrival slot and A^- and A^+ the number of arrivals during slot J respectively before and after the tagged customer has arrived. This approach, however, is incorrect. The reason is that the timer mechanism only matters after the queueing delay. Indeed, the timer is started only when the server becomes available and finds less than l customers. Shifting the postponing delay to the same instant as the start of the queueing delay and making use of relation (5.35) would assume that the timer is already counting during the queueing delay. Therefore, we have to resort to another approach.

Let $\hat{W}_2$ represent the postponing delay that starts at the same instant as the queueing delay. We however disconnect the postponing delay $\hat{W}_2$ from the timer mechanism β: $\hat{W}_2$ represents the number of slots until the batch containing the tagged customer can be filled with at least l customers. Next, define Θ as the time period after the queueing delay W_1 until the server decides to start a new service anyway, assuming that always less than l customers are present. The random variable Θ is by definition geometrically distributed, with probability distribution

$$\Pr[\Theta = n] = \beta(1-\beta)^n \quad , \qquad n \geq 0 \quad ,$$

or, equivalently,

$$\Pr[\Theta > n] = (1-\beta)^{n+1} \quad , \qquad n \geq 0 \quad .$$

On account of these definitions, we have the following relation between W, W_1, $\hat{W}_2$ and Θ (see also Fig. 5.23):

$$W = \begin{cases} W_1 & \text{if } W_1 \geq \hat{W}_2 \ , \\ \hat{W}_2 & \text{if } W_1 < \hat{W}_2 \text{ and } \hat{W}_2 \leq W_1 + \Theta \ , \\ W_1 + \Theta & \text{if } \hat{W}_2 > W_1 + \Theta \ . \end{cases}$$

This relation can be rewritten as:

$$W = \max\left(W_1, \min\left(\hat{W}_2, W_1 + \Theta\right)\right) \quad .$$

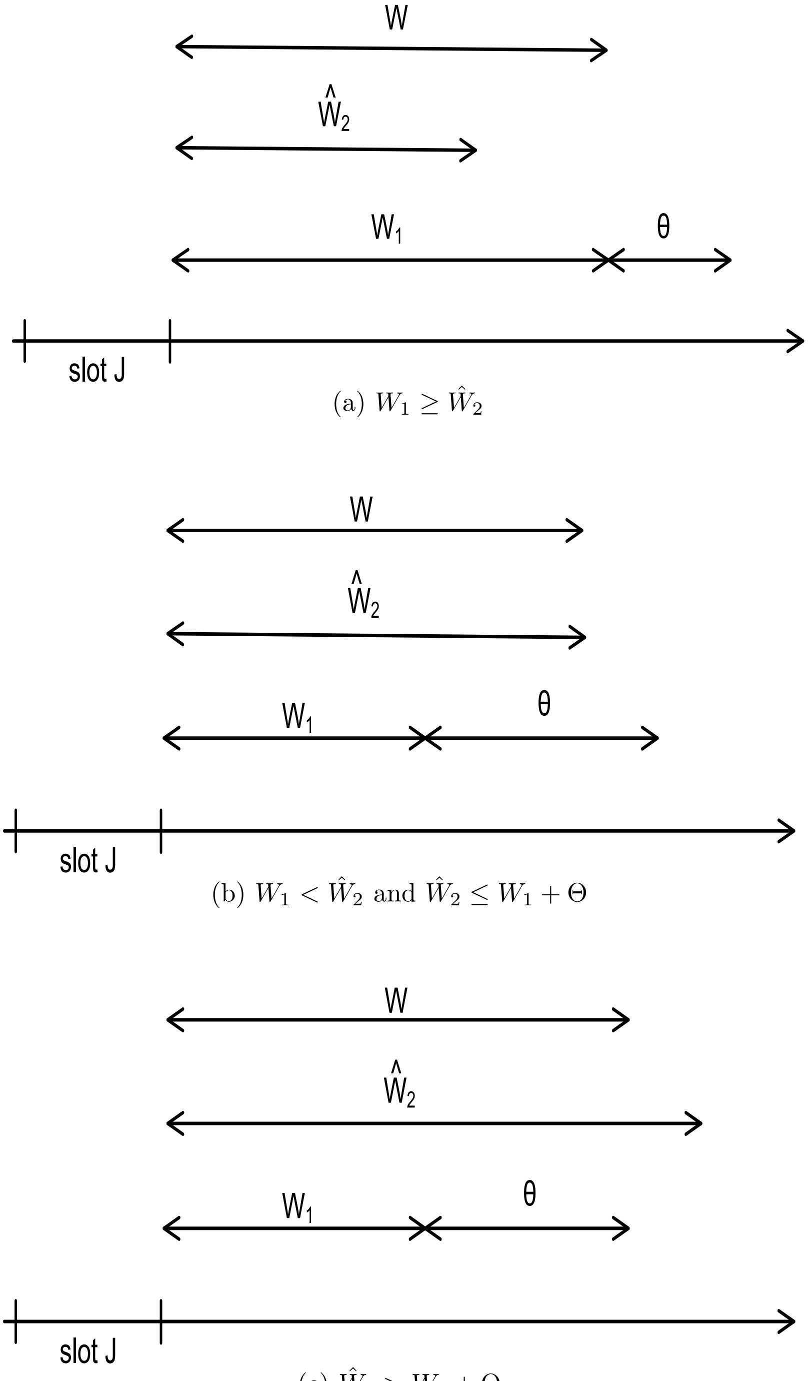
As a result, we find that

$$\begin{aligned}
\Pr[W > w] =& \Pr\left[\max\left(W_1, \min\left(\hat{W}_2, W_1 + \Theta\right)\right) > w\right] \\
=& \Pr\left[W_1 > w \vee \min\left(\hat{W}_2, W_1 + \Theta\right) > w\right] \\
=& \Pr[W_1 > w] + \Pr\left[\min\left(\hat{W}_2, W_1 + \Theta\right) > w\right] \\
& - \Pr\left[W_1 > w \wedge \min\left(\hat{W}_2, W_1 + \Theta\right) > w\right] \\
=& \Pr[W_1 > w] + \Pr\left[\hat{W}_2 > w \wedge W_1 + \Theta > w\right] \\
& - \Pr\left[W_1 > w \wedge \hat{W}_2 > w \wedge W_1 + \Theta > w\right] \\
=& \Pr[W_1 > w] + \Pr\left[W_1 + \Theta > w \wedge \hat{W}_2 > w\right] - \Pr\left[W_1 > w \wedge \hat{W}_2 > w\right] \quad .
\end{aligned}$$

As was the case in the previous sections, calculation of joint probabilities is difficult. Therefore, we resort to an approximation: we assume that $\hat{W}_2$ is independent of W_1 and thus of $W_1 + \Theta$ (because W_1 and $\hat{W}_2$ are independent of Θ), leading to the following expression:

$$\Pr[W > w] \approx \Pr[W_1 > w] + \Pr\left[\hat{W}_2 > w\right]\left\{\Pr[W_1 + \Theta > w] - \Pr[W_1 > w]\right\} \quad . \tag{5.36}$$

We now calculate $\Pr[W_1 > w]$, $\Pr[W_1 + \Theta > w]$ and $\Pr\left[\hat{W}_2 > w\right]$.



(a) $W_1 \geq \hat{W}_2$

(b) $W_1 < \hat{W}_2$ and $\hat{W}_2 \leq W_1 + \Theta$

(c) $\hat{W}_2 > W_1 + \Theta$

Figure 5.23: Illustration of relations between W, W_1, $\hat{W}_2$ and Θ

5.4.1 Calculation of $\Pr[W_1 > w]$

It was established in section 4.3 formula (4.18) that the PGF $W_1(z)$ of W_1 reads

$$
\begin{aligned}
W_1(z) = P(z,1) = \frac{T_c(z)-1}{c\lambda T_c(z)} \sum_{i=0}^{c-1} & \frac{A\left(T_c(z)^{1/c}\varepsilon_i\right)-1}{\left(T_c(z)^{1/c}\varepsilon_i-1\right)^2} \frac{T_c(z)^{1/c}\varepsilon_i}{z-A\left(T_c(z)^{1/c}\varepsilon_i\right)} \\
& \Bigg\{(z-1)(1-\beta)\sum_{n=0}^{l-1} d(n)\left(T_c(z)^{1/c}\varepsilon_i\right)^n \\
& +\beta\sum_{n=0}^{l-1} d(n)\left[T_n(z)-\left(T_c(z)^{1/c}\varepsilon_i\right)^n\right] \\
& +\sum_{n=l}^{c-1} d(n)\left[T_n(z)-\left(T_c(z)^{1/c}\varepsilon_i\right)^n\right]\Bigg\} \ . \qquad (5.37)
\end{aligned}
$$

In the sequel, we compute $\Pr[W_1 > w]$ by applying Darboux's theorem on (5.37). Therefore, it is required that the dominant singularities of $W_1(z)$ are known. Analogously as for the intermediate model, the dominant singularities are difficult to locate as compared to the basic model with single-slot service times, $\beta = 0$ and $l = c$. Indeed, the singularities of $W_1(z)$ might consist of zeroes of $T_c(z)$ outside the closed complex unit disk, zeroes of $T_c(z)^{1/c}\varepsilon_i - 1$ outside the closed complex unit disk, zeroes of $z - A\left(T_c(z)^{1/c}\varepsilon_i\right)$ outside the complex unit disk, possible singularities of $T_n(z)$, for $0 \le n \le c$, and possible singularities of $A(T_c(z)^{1/c}\varepsilon_i)$. Along the same lines as for the intermediate model we establish several theorems that play a crucial role in locating the dominant singularities.

Theorem 5. *Zeroes of $T_c(z)$ in the denominator of $W_1(z)$ produce no poles.*

The proof of this theorem is completely analogous as for theorem 2 (page 97).

Theorem 6. *The factor $(T_c(z)^{1/c}\varepsilon_i - 1)^2$ in the denominator produces no poles for $0 \le i \le c-1$.*

The proof of this theorem is completely analogous as for theorem 3 (page 98).

Lemma 2. *Assumptions 1-5 imply that $z^c - T_c(A(z))$ has exactly one zero in the interval $]1, \Re[$, where $\Re$ was defined in section 1.6 as $\min\{\Re_n : 0 \le n \le c\}$, with $\Re_n$ the radius of convergence of $T_n(A(z))$. In addition, the zero has multiplicity one and $z^c - T_c(A(z))$ contains no other zeroes with a modulus larger than one and smaller than or equal to this real zero.*

Proof. This lemma has been proved in [106]. □

Let us denote the only zero of $z^c - T_c(A(z))$ in the interval $]1, \Re[$ by $\tilde{z}$. Since $\tilde{z} < \Re \le \Re_A$, the following definition makes sense:

$$\hat{z} \triangleq A(\tilde{z}) \ .$$

It holds that $\hat{z} \in \mathbb{R}$ and $\hat{z} > 1$, since $A(1) = 1$, $\tilde{z} > 1$ and the PGF $A(z)$ being a real-valued and monotonically increasing function within $[1, \Re_A[$. In addition, as $\tilde{z} < \Re \leq \Re_c$, it follows that $\hat{z} = A(\tilde{z}) < \Re_T \leq \Re_{T_c}$, whereby $\Re_T$ was previously defined as $\min\{\Re_{T_n} : 0 \leq n \leq c\}$, with $\Re_{T_n}$ the radius of convergence of $T_n(z)$.

Theorem 7. *Assumptions 1-5 imply that*

1. $T_c(\hat{z})^{1/c} < \Re_A$ *and* $\hat{z}$ *is a zero of* $z - A(T_c(z)^{1/c})$*;*
2. *the equations* $z - A(T_c(z)^{1/c}\varepsilon_i)$, $0 \leq i \leq c-1$ *contain no other zeroes with a modulus larger than one and smaller than or equal to* $\hat{z}$*;*
3. $\hat{z}$ *is a zero of multiplicity one.*

The proof of this theorem is completely analogous as for theorem 4 (page 98).

Summarizing the theorems and taking into account that $\hat{z} < \Re_T$, $W_1(z)$ has one dominant singularity, being a pole $\hat{z}$. This dominant pole is a real number larger than one, is a zero of $z - A(T_c(z)^{1/c})$, has multiplicity one and is equal to $A(\tilde{z})$, with $\tilde{z}$ the only zero in $]1, \Re[$ of $z^c - T_c(A(z))$. As $\tilde{z} \in \mathbb{R}$, it can be easily determined numerically, for instance with the bisection or the Newton-Raphson method.
Taking these findings into account, we obtain that $W_1(z)$ is in the neighborhood of $\hat{z}$ proportional to

$$W_1(z) \sim \frac{G(z)}{z - A\left(T_c(z)^{1/c}\right)} \quad ,$$

with

$$\begin{aligned} G(z) = & \frac{T_c(z)-1}{c\lambda T_c(z)} \frac{A\left(T_c(z)^{1/c}\right)-1}{\left(T_c(z)^{1/c}-1\right)^2} T_c(z)^{1/c} \left\{ (z-1)(1-\beta) \sum_{n=0}^{l-1} d(n) T_c(z)^{n/c} \right. \\ & \left. + \beta \sum_{n=0}^{l-1} d(n) \left[T_n(z) - T_c(z)^{n/c}\right] + \sum_{n=l}^{c-1} d(n) \left[T_n(z) - T_c(z)^{n/c}\right] \right\} \quad . \end{aligned}$$

Hence, application of formula (1.4) of Darboux's theorem yields

$$\Pr\left[W_1 > w\right] \approx \frac{\hat{z}^{-(w+1)}}{1-\hat{z}} \frac{cG(\hat{z})}{c - A'\left(T_c(\hat{z})^{1/c}\right) T_c(\hat{z})^{\frac{1}{c}-1} T_c'(\hat{z})} \quad . \tag{5.38}$$

5.4.2 Calculation of $\Pr\left[W_1 + \Theta > w\right]$

It seems evident to calculate this probability by relating it with $\Pr\left[W_1 > w\right]$:

$$\Pr\left[W_1 + \Theta > w\right] = \sum_{t=0}^{w} \Pr\left[\Theta = t\right] \Pr\left[W_1 > w - t\right] + \sum_{t=w+1}^{\infty} \Pr\left[\Theta = t\right] \quad .$$

In this expression we could then use approximation (5.38) for $\Pr\left[W_1 > w\right]$. This approach however, might lead to inaccurate results as for t approaching w, formula (5.38) for $\Pr\left[W_1 > w - t\right]$ can be inaccurate because $w - t$ is very small. Let us therefore consider the PGF corresponding to $W_1 + \Theta$, which is, because Θ is independent of W_1, equal to $W_1(z)\Theta(z)$, with

$$\Theta(z) \triangleq \mathrm{E}\left[z^{\Theta}\right] = \sum_{n=0}^{\infty} (1-\beta)^n \beta z^n = \frac{\beta}{1-(1-\beta)z} \quad .$$

Note that the dominant singularity of $\Theta(z)$, say z^*, is equal to $1/(1-\beta)$. The dominant singularity of $W_1(z)\Theta(z)$ is thus either equal to z^* or $\hat{z}$, depending on which is the smallest. We thus have to consider three scenarios.

$z^* < \hat{z}$
The dominant pole thus equals z^* and it has multiplicity one. Hence, application of formula (1.4) of Darboux's theorem yields

$$\Pr\left[W_1 + \Theta > w\right] \approx \frac{(z^*)^{-(w+1)}}{1-z^*}\frac{-\beta}{1-\beta}W_1(z^*) = (1-\beta)^{w+1}W_1\left(\frac{1}{1-\beta}\right) \ .$$

$z^* > \hat{z}$
In this case, $\hat{z}$ is the dominant pole. As it has multiplicity one, the approximation is obtained by applying formula (1.4) of Darboux's theorem, leading to

$$\Pr\left[W_1 + \Theta > w\right] \approx \frac{\hat{z}^{-(w+1)}}{1-\hat{z}}\frac{\beta}{1-(1-\beta)\hat{z}}\frac{cG(\hat{z})}{c - A'\left(T_c(\hat{z})^{1/c}\right)T_c(\hat{z})^{\frac{1}{c}-1}T_c'(\hat{z})} \ .$$

$z^* = \hat{z}$
In this case, $\hat{z} = z^*$ is the dominant pole and it has multiplicity two. We thus find that $W_1(z)\Theta(z)$ is proportional to

$$W_1(z)\Theta(z) \sim G(z)\beta\frac{1-(1-\beta)z}{z - A\left(T_c(z)^{1/c}\right)}\left[1 - \frac{z}{1/(1-\beta)}\right]^{-2} \ . \tag{5.39}$$

As a result, $\Pr\left[W_1 + \Theta > w\right]$ can be deduced by application of formula (1.3) of Darboux's theorem on (5.39), resulting in

$$\Pr\left[W_1 + \Theta > w\right] \approx \frac{(z^*)^{-w}}{z^*-1}G(z^*)\beta\frac{-(1-\beta)c}{c - A'(T_c(z^*)^{1/c})T_c(z^*)^{1/c-1}T_c'(z^*)}w \ . \tag{5.40}$$

Remark 26. *When $z^* \neq \hat{z}$ but $z^* \approx \hat{z}$, it is better to take into account both contributions of z^* and $\hat{z}$ in* $\Pr\left[W_1 > w\right]$ *(see e.g. [107]), leading to:*

$$\begin{aligned}\Pr\left[W_1 + \Theta > w\right] \approx &(1-\beta)^{w+1}W_1\left(\frac{1}{1-\beta}\right)\\ &+ \frac{\hat{z}^{-(w+1)}}{1-\hat{z}}\frac{\beta}{1-(1-\beta)\hat{z}}\frac{cG(\hat{z})}{c - A'\left(T_c(\hat{z})^{1/c}\right)T_c(\hat{z})^{\frac{1}{c}-1}T_c'(\hat{z})} \ .\end{aligned} \tag{5.41}$$

We adopt this approach in the numerical examples in section 5.4.4.

5.4.3 Calculation of $\Pr\left[\hat{W}_2 > w\right]$

In order to calculate $\Pr\left[\hat{W}_2 > w\right]$, we start from the following relation:

$$\Pr\left[\hat{W}_2 > w\right] = \Pr\left[\left(\left[Q_J + A^-\right] \bmod c\right) + 1 + A^+ + \sum_{i=1}^{w} A_{J+i} < l\right] \ . \tag{5.42}$$

Indeed, $\hat{W}_2$ is larger than w if the sum of (a) the number of previously arrived customers that are served in the same batch as the tagged customer

($[Q_J + A^-] \bmod c$), (b) the tagged customer, (c) the number of customer arrivals during slot J after the tagged customer (A^+) and (d) the number of arrivals during the sequence of w slots following slot J ($\sum_{i=1}^{w} A_{J+i}$), is smaller than the threshold l.

As a next step, we transform expression (5.42) by means of the probability generating property of PGFs. Since A^+ and A^- are correlated, but independent of the other discrete random variables that appear in (5.42) (due to the IID nature of the arrivals), we first compute $\mathrm{E}\left[x^{([Q_J+A^-] \bmod c)}\, x^{A^+}\right]$. Analogously as in sections 5.2 (formula (5.10)) and 5.3 (formula (5.26)), we find that

$$\mathrm{E}\left[x^{([Q_J+A^-] \bmod c)}\, x^{A^+}\right] = \frac{x^c-1}{c(x-1)} \sum_{i=0}^{c-1} Q(\varepsilon_i) \frac{A(\varepsilon_i)-A(x)}{\lambda(\varepsilon_i - x)} \frac{\varepsilon_i(x-1)}{x-\varepsilon_i} \quad . \tag{5.43}$$

As $Q(\varepsilon_0) = Q(1) = 1$ and on account of expression (2.17) for $Q(z)$, we obtain

$$Q(\varepsilon_i) = \frac{\beta \sum_{n=0}^{l-1} d(n)(\varepsilon_i^n - 1) + \sum_{n=l}^{c-1} d(n)(\varepsilon_i^n - 1)}{A(\varepsilon_i) - 1} \quad , \qquad 1 \le i \le c-1 \quad . \tag{5.44}$$

The combination of (5.43) and (5.44) produces

$$\mathrm{E}\left[x^{([Q_J+A^-] \bmod c)}\, x^{A^+}\right] = \frac{x^c-1}{c(x-1)} g(x) \quad , \tag{5.45}$$

with

$$\begin{aligned} g(x) = {} & \frac{1-A(x)}{\lambda(1-x)} \\ & + \sum_{i=1}^{c-1} \frac{A(\varepsilon_i)-A(x)}{\lambda(\varepsilon_i - x)} \frac{\varepsilon_i(x-1)}{x-\varepsilon_i} \frac{\beta \sum_{n=0}^{l-1} d(n)(\varepsilon_i^n - 1) + \sum_{n=l}^{c-1} d(n)(\varepsilon_i^n - 1)}{A(\varepsilon_i) - 1} \quad , \end{aligned}$$

whereby the first term in the right-hand-side represents $i = 0$. The combination of (5.42), (5.45) and the probability generating property of PGFs yields

$$\begin{aligned} \Pr\left[\hat{W}_2 > w\right] &= \sum_{m=0}^{l-1} \frac{1}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} \mathrm{E}\left[x^{([Q_J+A^-] \bmod c)+1+A^+ + \sum_{i=1}^{w} A_{J+i}}\right]\Bigg|_{x=0} \\ &= \sum_{m=1}^{l-1} \frac{1}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} x A(x)^w \frac{x^c-1}{c(x-1)} g(x) \Bigg|_{x=0} \quad . \end{aligned}$$

On account of Leibniz's rule for the derivative of a product, this can, analogously as for the intermediate model in section 5.3 (formula (5.28)), be transformed into

$$\Pr\left[\hat{W}_2 > w\right] = \frac{1}{c} \sum_{m=0}^{l-2} \frac{l-1-m}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} A(x)^w g(x) \Bigg|_{x=0} \quad . \tag{5.46}$$

Formula (5.46) can be implemented in a mathematical program such as matlab. This procedure suffers from the drawback that high-order derivatives may have to be computed, which causes a considerable reduction in speed and even is infeasible if l and c are quite large. Therefore, we deduce an approximation for $\Pr\left[\hat{W}_2 > w\right]$, whereby no derivatives have to be calculated.

As the methodology runs completely along the same lines as in section 5.3, we immediately mention the approximation formula that we eventually obtain:

$$\Pr\left[\hat{W}_2 > w\right] \approx \frac{g(0)}{c} \sum_{m=0}^{\min(l-2,w)} A'(0)^m (l-1-m) \binom{w}{m} A(0)^{w-m} \quad . \tag{5.47}$$

Note that for large w, formula (5.47) becomes a sum from 0 to $l-2$. We further point out that the binomial coefficient causes no difficulties, since efficient routines exist to calculate them, even for large w.

Remark 27. *When $l=1$, $\Pr\left[\hat{W}_2 > w\right]$ equals 0. Of course, when the service threshold equals 1, a present customer cannot suffer a delay due to postponing service until more customers have arrived.*

Remark 28. *Note again that this approach is not suited for cases whereby $A^{'}(0)=0$, as only the term corresponding to $m=0$ in (5.47) differs from 0 and this term is the one with the smallest power of $1-zA(0)$ in the denominator of $\hat{W}_2(z)$.*

Remark 29. *When $T_n(z) = T_c(z)$ $\forall n$ and $\beta = 0$, (5.47) reduces to the corresponding expression (5.34) for the intermediate model.*

5.4.4 Evaluation of approximation formulas

In this section, we evaluate the accuracy of our approach. First, we study formula (5.38) for $\Pr[W_1 > w]$. Then, we focus on expressions (5.40)-(5.41) for $\Pr[W_1 + \Theta > w]$, next on approximation (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ and finally expression (5.36) for $\Pr[W > w]$ is covered.

In Figures 5.24-5.26, we have depicted approximation (5.38) as well as simulated values[3] for $\Pr[W_1 > w]$ versus w for various combinations of distributions for the number of customer arrivals (Poisson $A(z) = e^{\lambda(z-1)}$; Geometric $A(z) = 1/(1+\lambda-\lambda z)$; C-center $A(z) = 1-\lambda/c+\lambda/(2c)(z^{c-1}+z^{c+1})$) and service times (Geometric $T_n(z) = z/[\mathrm{E}[T_n] + (1-\mathrm{E}[T_n])z]$ with $\mathrm{E}[T_n] = 8+0.2n$ or $\mathrm{E}[T_n] = 6+0.4n$; 25 $T_c(z) = (25-\mathrm{E}[T_c])z/24+(\mathrm{E}[T_c]-1)z^{25}/24$ with $\mathrm{E}[T_n] =$ 5 or 10) and several server capacities c, service thresholds l, timer parameters β and loads ρ. We observe that the approximation is, as in the previous sections, accurate. However, in the case $\rho = 0.3$ in Fig. 5.24 (a) it is less effective. In order to understand why, we have reported in Table 5.1 the dominant pole $\hat{z}$ of $W_1(z)$ versus the load and the singularities of $T_n(z)$ (in case of geometrically distributed service times, $\gamma_n \triangleq \mathrm{E}[T_n]/[\mathrm{E}[T_n]-1]$ is the singularity of $T_n(z)$) and we have depicted in Fig. 5.27 the approximation and the simulated values of $\Pr[W_1 > 75]$ versus the load for several expressions of $A(z)$ and $T_n(z)$. We notice in Table 5.1 that, in case of Poisson and geometric arrivals, the smaller the load (and thus the smaller the mean arrival rate λ), the more $\hat{z}$ approaches to the singularity γ_c, which is the reason why the approximation becomes inaccurate for small load in these cases. This anomaly is not specific for our model

[3] Only for $\Pr\left[\hat{W}_2 > w\right]$ we have an exact formula at our disposal. For the other tail probabilities ($\Pr[W_1 > w]$, $\Pr[W_1 + \Theta > w]$ and $\Pr[W > w]$) throughout this section, we have therefore depicted the 95% confidence intervals resulting from 10 Monte Carlo simulations whereby each simulation generates W_1, $W_1 + \Theta$ and W for 10^8 customers.

but is inherent to approximations based on dominant singularities in general. In case of c-centered arrivals, $\hat{z}$ approaches γ_n a lot slower, which entails a much better accuracy of the approximation. Fig. 5.27 also exhibits that the approximation is precise in case of Poisson arrivals and services of either 1 or 25 slots with mean value 5 (regardless of the number of customers in the served batch). The reason is that $T_n(z)$ has no singularities in this case. In general, we can conclude that the approximation is accurate except when the load is small in combination with $T_n(z)$ (and/or $A(z)$) having singularities. In such situations, it is possible to enhance the approximation by adopting an ad hoc approach whereby the contributions of the other singularity(ies) nearby $\hat{z}$ is (are) also incorporated (see e.g. [107]).

Remark 30. *In section 5.3.3, we have also considered an example with Poisson arrivals and geometric service times. There however, the approximation for* $\Pr[W_1 > w]$ *did not suffer as much as here from the singularity of the PGF of the service times. The reason for this is that we here deal with several PGFs of the service times depending on the number of customers in the served batch and that the dominant pole* $\hat{z}$ *of* $W_1(z)$ *also approaches the singularities of* $T_{c-1}(z)$*,* $T_{c-2}(z)$*, et cetera when the load becomes very small. It has been shown that approximations based on dominant singularities become less accurate the more other singularities approach the dominant singularity(ies) (see e.g. [107]).*

Next, we evaluate approximations (5.40)-(5.41) for $\Pr[W_1 + \Theta > w]$ in Figures 5.28-5.30. We observe that these are accurate. Indeed, for larger loads, W_1 is dominant in $W_1 + \Theta$ and the approximation for $\Pr[W_1 > w]$ is precise for larger loads, whereas for smaller loads, W_1 becomes small, so that Θ determines the behaviour of $W_1 + \Theta$ and the formula for $\Pr[\Theta > w]$ is exact (due to its geometric distribution). In some special cases however, the approximation might become inaccurate for smaller loads. Consider for instance the system with Poisson arrivals, geometric service times with $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$ and $\beta = 0.2$. In this situation, the singularity of $\Theta(z)$ equals $1/(1-\beta) = 1.25$ and the singularity γ_c of $T_c(z)$ equals $1.11\ldots$ (see Table 5.1). We thus have that $1/(1-\beta)$ is, regardless of the load, larger than $\hat{z}$ (because $\hat{z} < \gamma_c$), which means that $\hat{z}$ is always the dominant pole of $W_1(z)\Theta(z)$, thus that W_1 even dominates $W_1 + \Theta$ for smaller loads, which results in an inaccurate approximation (see Fig. 5.31). When $\mathrm{E}[T_n]$ is equal to $3 + 0.1n$, γ_c is equal to $1.333\ldots$, so that Θ will again dominate for smaller loads, which thus leads to a good approximation (see Fig. 5.32). Hence, approximations (5.40)-(5.41) for $\Pr[W_1 + \Theta > w]$ are accurate, except for special cases whereby $1/(1-\beta)$ is always larger than $\hat{z}$, which leads to bad results for small loads. Again, this can be resolved by following an ad hoc approach in such situations.

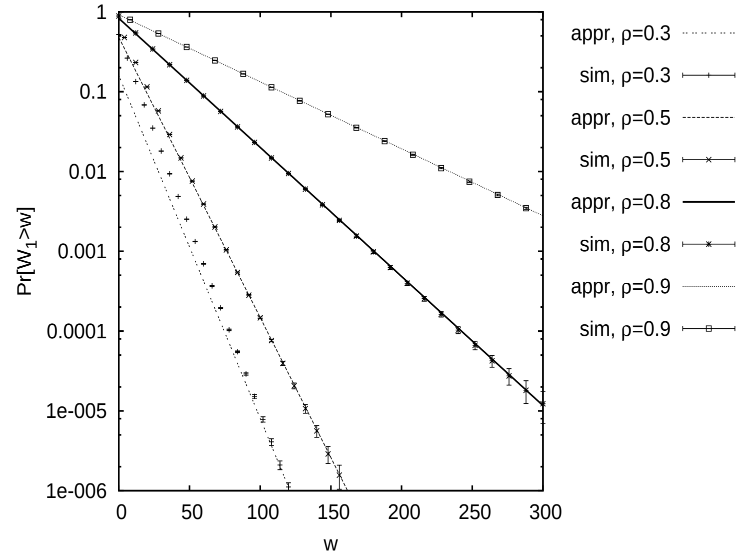
Next, approximation (5.47) and exact formula (5.46) for $\Pr\left[\hat{W}_2 > w\right]$ are depicted versus w in Figures 5.33-5.35 for various settings of the system parameters. Along the same lines as for the basic model, we observe that the approximation is accurate for larger values of w, except when $A^{'}(0) = 0$, and thus that it is extremely suited for quickly assessing the order of magnitude of $\Pr\left[\hat{W}_2 > w\right]$.

Finally, we investigate approximation (5.36) for $\Pr[W > w]$ in Figures 5.36-5.40. We observe, in general, that the approximation is accurate, except for values of the load between, roughly speaking, 0.15 and 0.35, where it is less precise but still acceptable for the purpose of assessing the order of magnitude of $\Pr[W > w]$. In order to explain this issue, all probabilities (queueing, postponing and total) are depicted versus the load in Figures 5.41-5.42. The approximation being extremely accurate for larger loads follows from $\Pr[W_1 > w]$ then clearly dominating in (5.36) and the approximation of $\Pr[W_1 > w]$ being outstanding in this area. For "medium" values of the load, $\Pr[W_1 > w] \approx \Pr\left[\hat{W}_2 > w\right]$, so that both play a role in (5.36). In this area however, the approximation for $\Pr[W_1 > w]$ is for geometric service times combined with Poisson arrivals or geometric arrivals not excellent but still adequate. When the load is small, $\Pr[W_1 > w] << \Pr\left[\hat{W}_2 > w\right]$, so that $\Pr\left[\hat{W}_2 > w\right]$ and $\Pr[W_1 + \Theta > w]$ dominate in (5.36). In addition, as the load is small, it generally holds that $\Pr[W_1 + \Theta > w] \approx \Pr[\Theta > w]$ (except in some special cases, of which we discuss one below). As the approximation for $\Pr\left[\hat{W}_2 > w\right]$ is precise and the formula for $\Pr[\Theta > w]$ is exact (it has a geometric distribution), the approximation is very accurate.

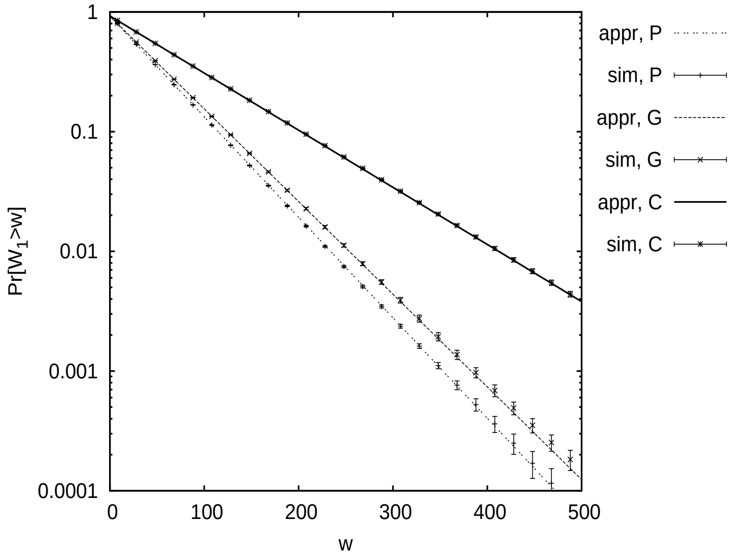
The c-centered arrivals, however, is an outsider: although the approximation for $\Pr[W_1 > w]$ is accurate in this case, the approximation for $\Pr[W > w]$ is inaccurate for smaller values of the load. This can be explained intuitively. Approximation (5.36) is mainly based on the assumption that $\Pr\left[\hat{W}_2 > w\right]$ is independent of $\Pr[W_1 > w]$, which is not a good assumption in this special case of c-centered arrivals and low load. Indeed, when W_1 is not equal to 0, this probably means, owing to the low load, that at slot mark $J + 1$ service is initiated of a batch consisting of some customers that arrived in slot J, but not containing the tagged customer itself. As a consequence, the probability is larger that there are at slot mark $J + 1$ not yet enough customers to fill the batch with the tagged customer sufficiently, which implies that $\hat{W}_2$ is likely to be large. When, on the other hand, $W_1 = 0$, $\hat{W}_2$ can only differ from zero when $c - 1$ customers arrive during slot J and if the system was empty at the beginning of that slot and if $l = c$. In other words, W_1 and $\hat{W}_2$ are strongly correlated.

Before closing this section, we again study our previously considered example where $\beta = 0.2$. When $\mathrm{E}[T_n] = 8 + 0.2n$ (Fig. 5.43), the approximation for $\Pr[W > w]$ is inaccurate for small loads, which is a direct result of W_1 dominating over Θ in this case, whereas when $\mathrm{E}[T_n] = 3 + 0.1n$ (Fig. 5.44), Θ again dominates over W_1 for small loads, which results in a good approximation.

Summarized, we feel that our approximation is **very useful for the purpose of assessing the order of magnitude of the customer delay**, except in some special cases.

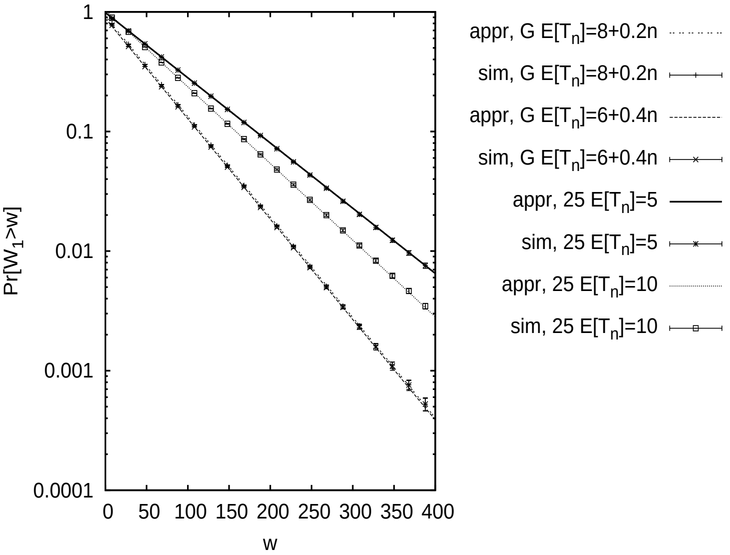


(a) several loads; $c = 10$, $l = 5$, $\beta = 0.05$, Poisson arrivals, geometric services
$\mathrm{E}[T_n] = 8 + 0.2n$

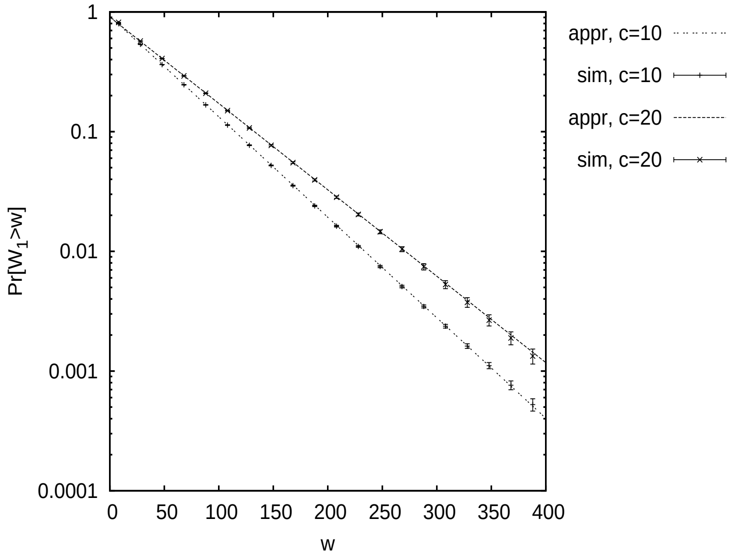


(b) several $A(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.9$, geometric services
$\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.24: Evaluation of approximation formula (5.38) for $\Pr[W_1 > w]$ (1)

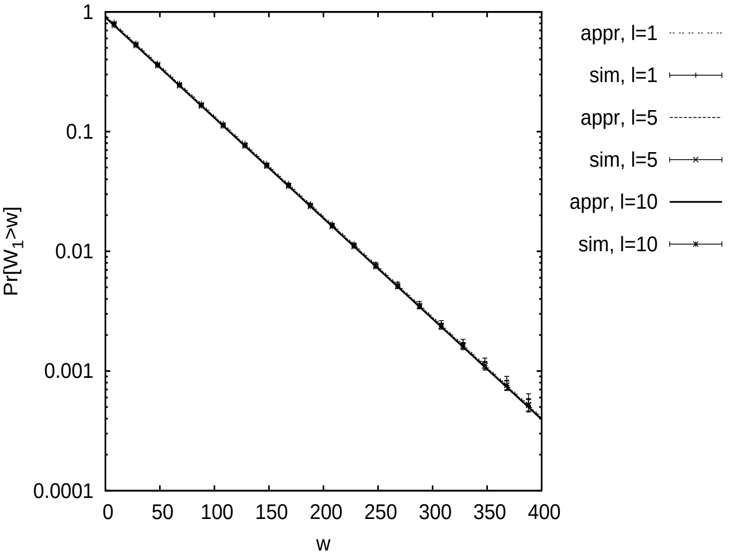


(c) several $T_n(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.9$, Poisson arrivals

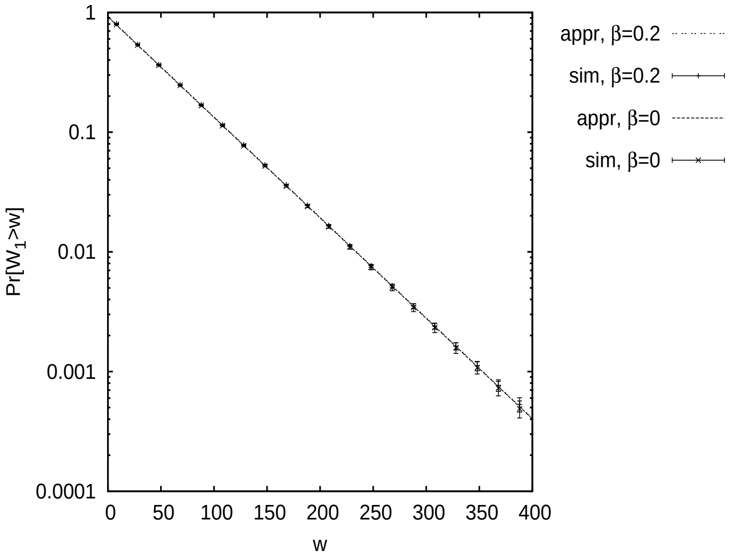


(d) several c's; $l = 5$, $\beta = 0.05$, $\rho = 0.9$, Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$

Figure 5.25: Evaluation of approximation formula (5.38) for $\Pr\left[W_1 > w\right]$ (2)



(e) several l's; $c = 10$, $\beta = 0.05$, $\rho = 0.9$, Poisson arrivals, geometric services mean $\mathrm{E}\,[T_n] = 8 + 0.2n$



(f) several β's; $c = 10$, $l = 5$, $\rho = 0.9$, Poisson arrivals, geometric services mean $\mathrm{E}\,[T_n] = 8 + 0.2n$

Figure 5.26: Evaluation of approximation formula (5.38) for $\Pr[W_1 > w]$ (3)

Table 5.1: Singularities γ_n of $T_n(z)$ when $T_n(z) = z/[\mathrm{E}[T_n] + (1 - \mathrm{E}[T_n])z]$, with $\mathrm{E}[T_n] = 8 + 0.2n$, and dominant pole $\hat{z}$ of $W_1(z)$ for several distributions of $A(z)$ and various values of the load ρ; $c = 10$

n	γ_n	ρ	$\hat{z}$ Poisson	$\hat{z}$ geometric	$\hat{z}$ c-centered
0	1.142857142857	0.9	1.019517053853	1.017984717482	1.011049828292
2	1.135135135135	0.7	1.055046820392	1.052088641246	1.033151838444
4	1.128205128205	0.5	1.084194576530	1.081656490042	1.055263630655
6	1.121951219512	0.3	1.104020768326	1.102948002054	1.077404642357
8	1.116279069767	0.1	1.111018197772	1.110989967865	1.099654929738
9	1.113636363636	0.05	1.111109639324	1.111109020127	1.105282554417
10	1.111111111111	0.01	1.111111111100	1.111111111106	1.109878832698

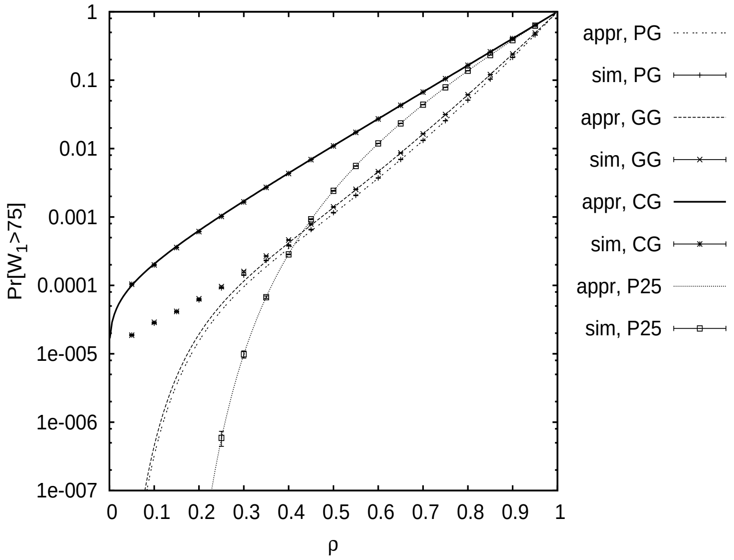
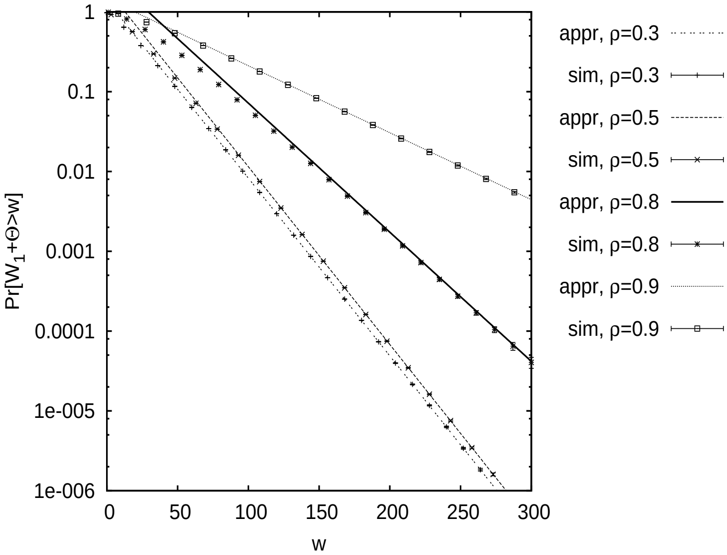
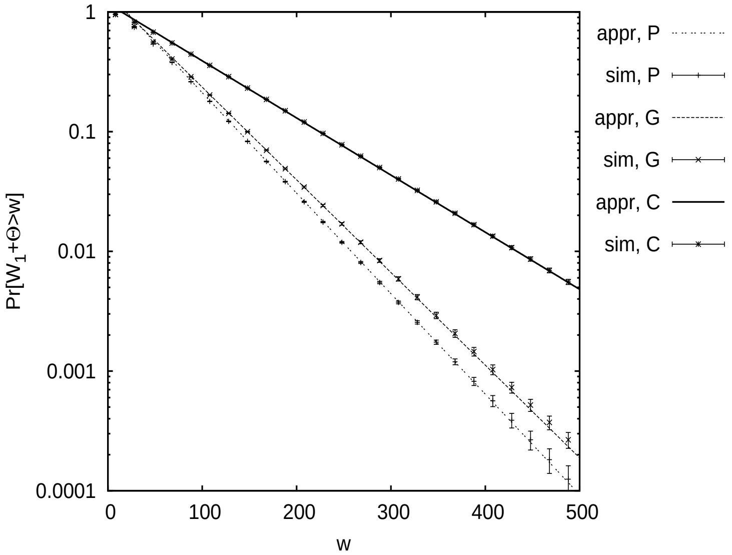


Figure 5.27: $\Pr[W_1 > 75]$ versus the load ρ for several combinations of $A(z)$ and $T_n(z)$ (PG Poisson arrivals geometric services $\mathrm{E}[T_n] = 8 + 0.2n$; GG geometric arrivals geometric services $\mathrm{E}[T_n] = 8 + 0.2n$; CG c-centered arrivals geometric services $\mathrm{E}[T_n] = 8 + 0.2n$; P25 Poisson arrivals 1 or 25 slots service $\mathrm{E}[T_n] = 5$); $c = 10$, $l = 5$, $\beta = 0.05$

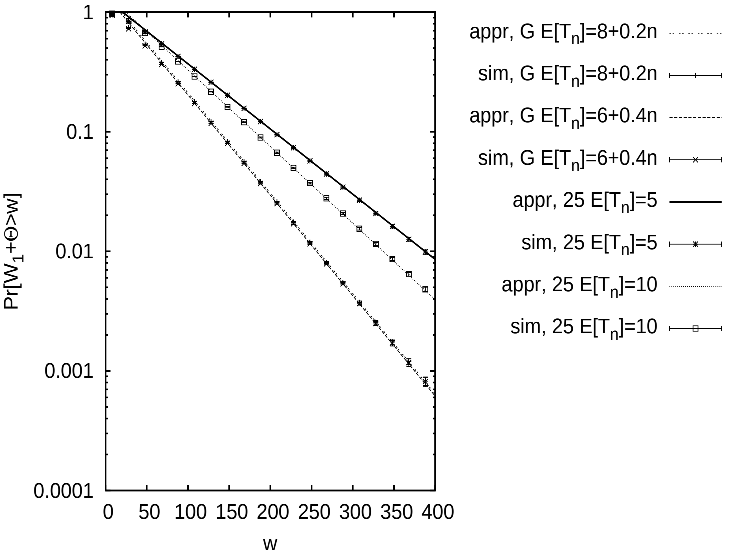


(a) several loads; $c = 10$, $l = 5$, $\beta = 0.05$, Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

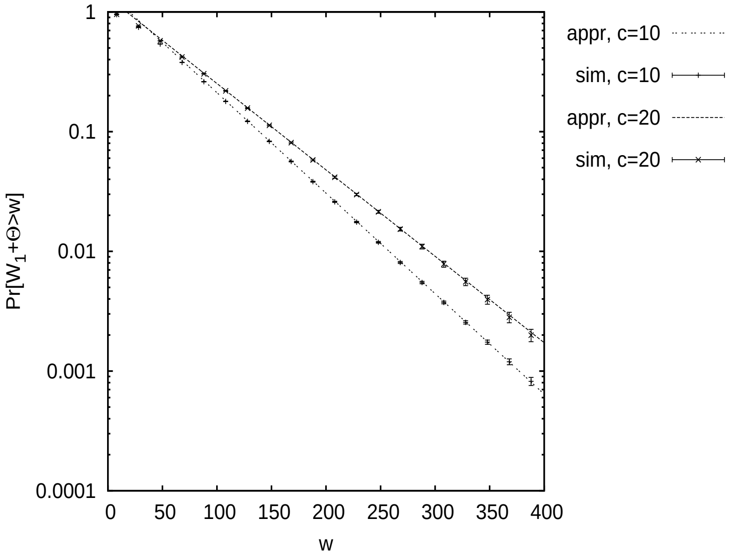


(b) several $A(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.9$, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.28: Evaluation of approximation formulas (5.40)-(5.41) for $\Pr[W_1 + \Theta > w]$ (1)

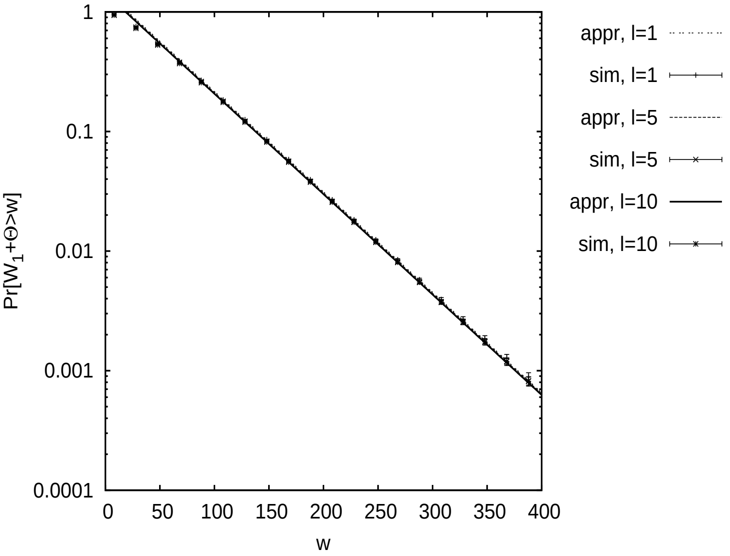


(c) several $T_n(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.9$, Poisson arrivals

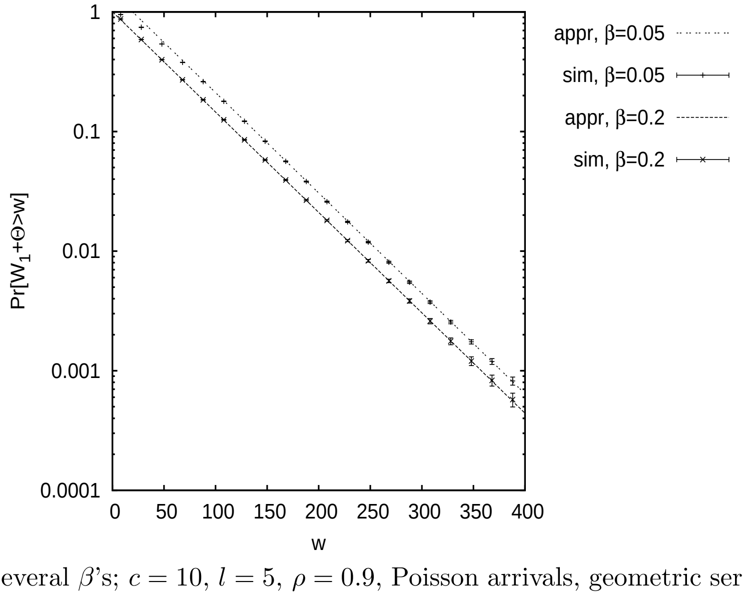


(d) several c's; $l = 5$, $\beta = 0.05$, $\rho = 0.9$, Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.29: Evaluation of approximation formulas (5.40)-(5.41) for $\Pr[W_1 + \Theta > w]$ (2)



(e) several l's; $c = 10$, $\beta = 0.05$, $\rho = 0.9$, Poisson arrivals, geometric services mean $\mathrm{E}\,[T_n] = 8 + 0.2n$



(f) several β's; $c = 10$, $l = 5$, $\rho = 0.9$, Poisson arrivals, geometric services mean $\mathrm{E}\,[T_n] = 8 + 0.2n$

Figure 5.30: Evaluation of approximation formulas (5.40)-(5.41) for $\Pr[W_1 + \Theta > w]$ (3)

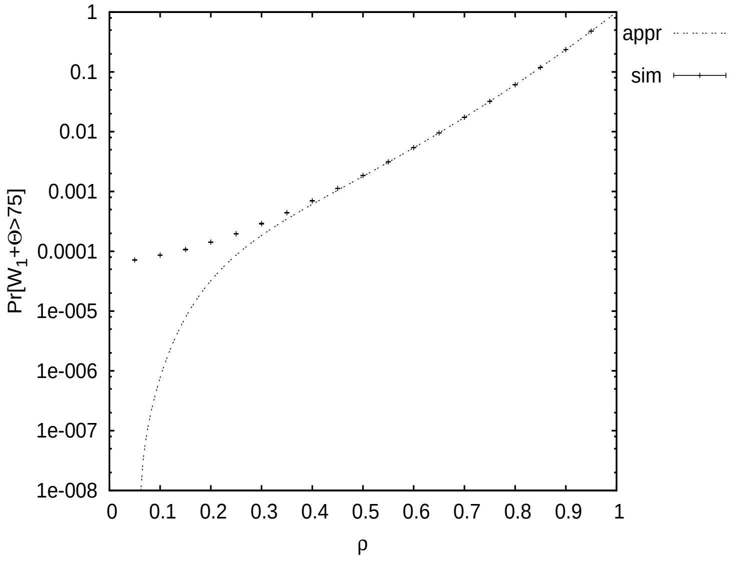


Figure 5.31: Evaluation of $\Pr\left[W_1 + \Theta > 75\right]$ versus the load ρ; Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$, $c = 10$, $l = 5$, $\beta = 0.2$

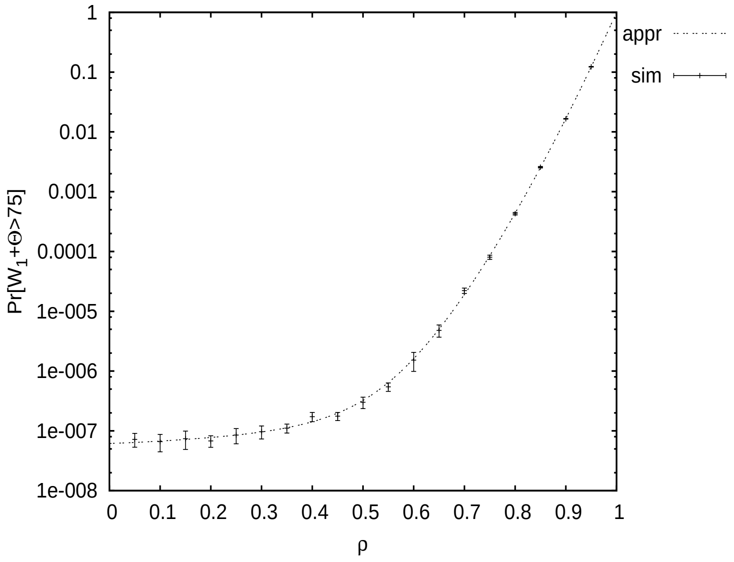
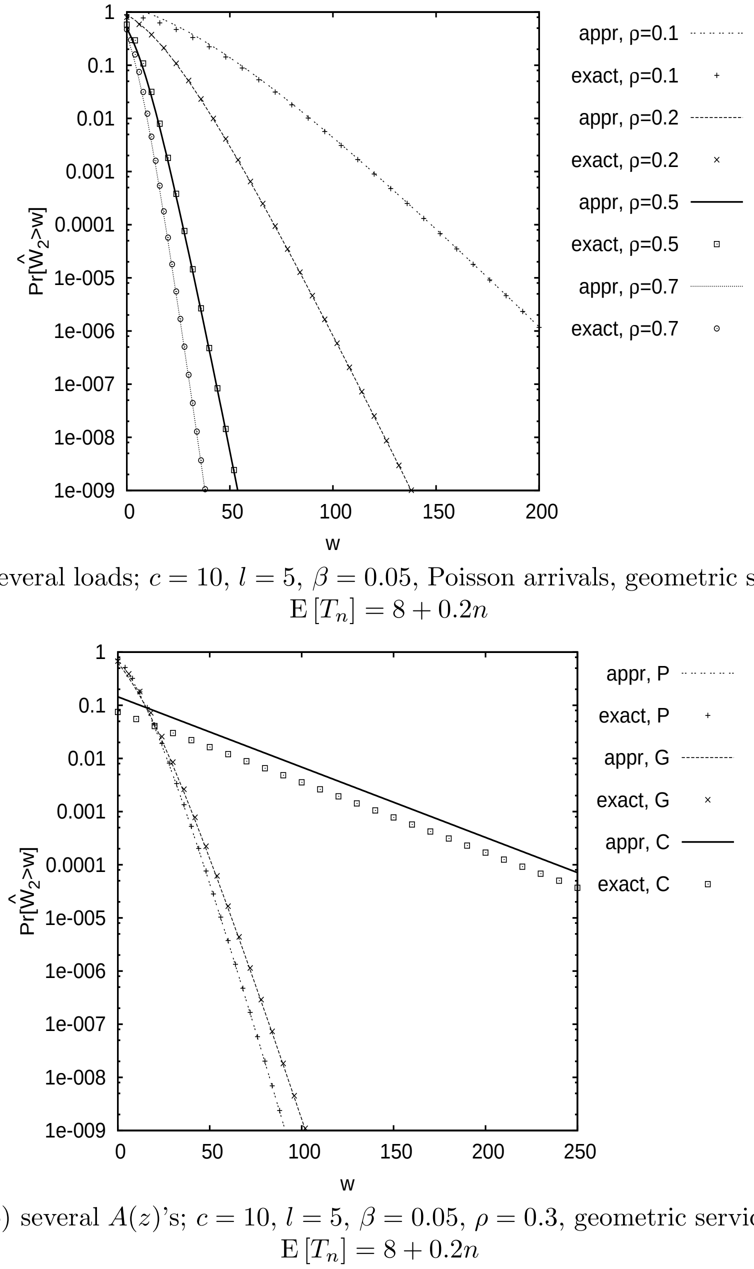


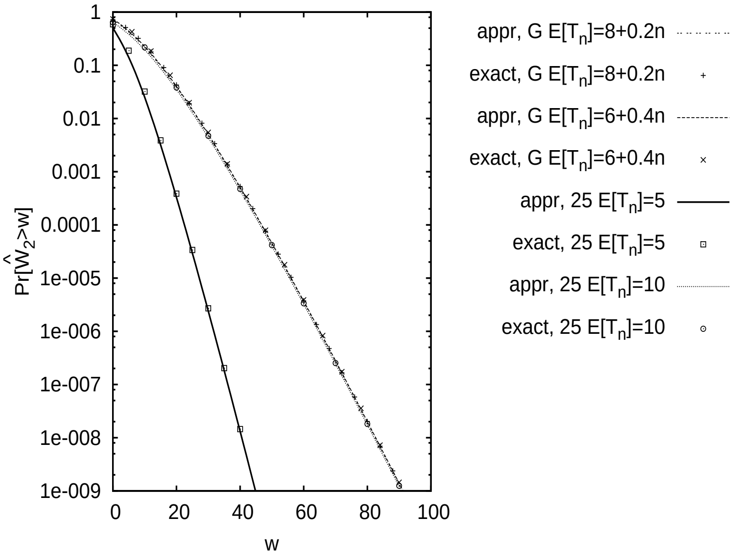
Figure 5.32: Evaluation of $\Pr\left[W_1 + \Theta > 75\right]$ versus the load ρ; Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 3 + 0.1n$, $c = 10$, $l = 5$, $\beta = 0.2$



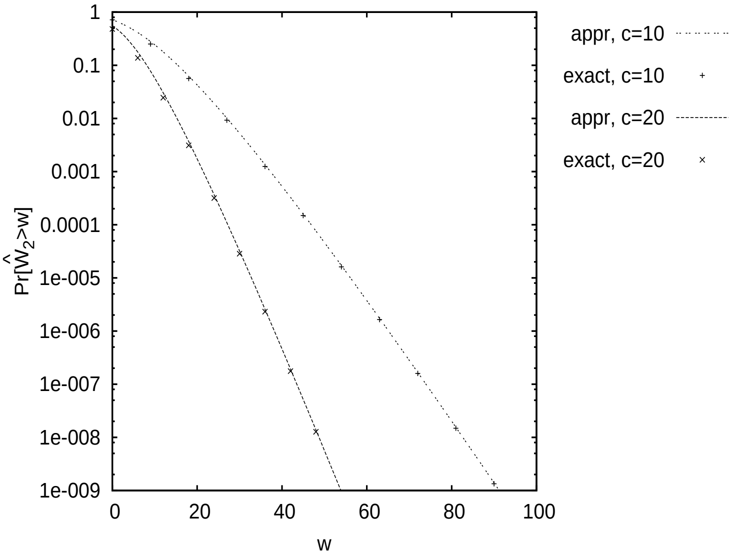
(a) several loads; $c = 10$, $l = 5$, $\beta = 0.05$, Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$

(b) several $A(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.3$, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$

Figure 5.33: Evaluation of approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ by comparing it with exact formula (5.46) (1)

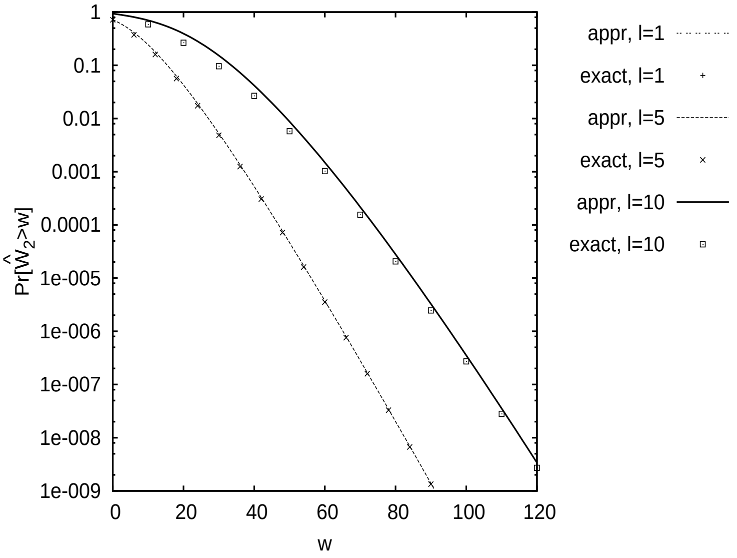


(c) several $T_n(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.3$, Poisson arrivals

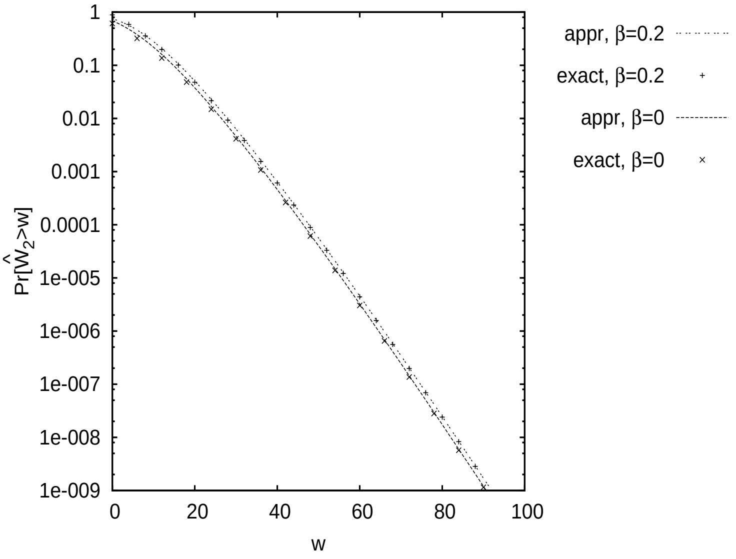


(d) several c's; $l = 5$, $\beta = 0.05$, $\rho = 0.3$, Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.34: Evaluation of approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ by comparing it with exact formula (5.46) (2)

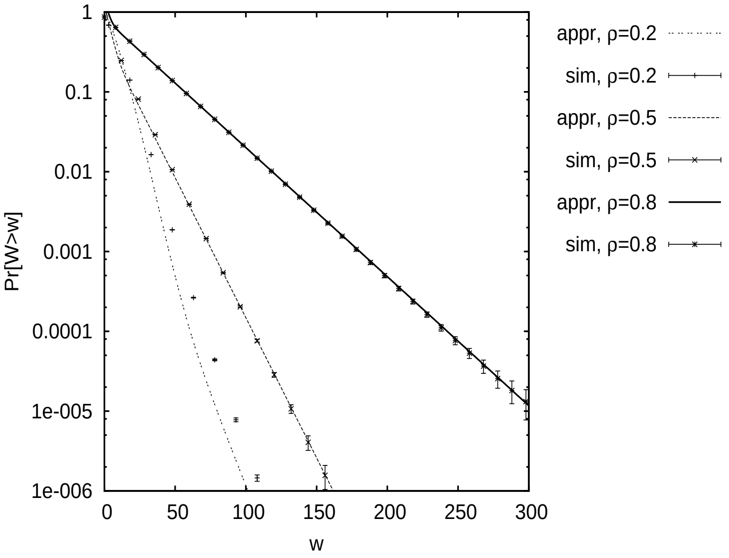


(e) several l's; $c = 10$, $\beta = 0.05$, $\rho = 0.3$, Poisson arrivals, geometric services mean $\mathrm{E}[T_n] = 8 + 0.2n$

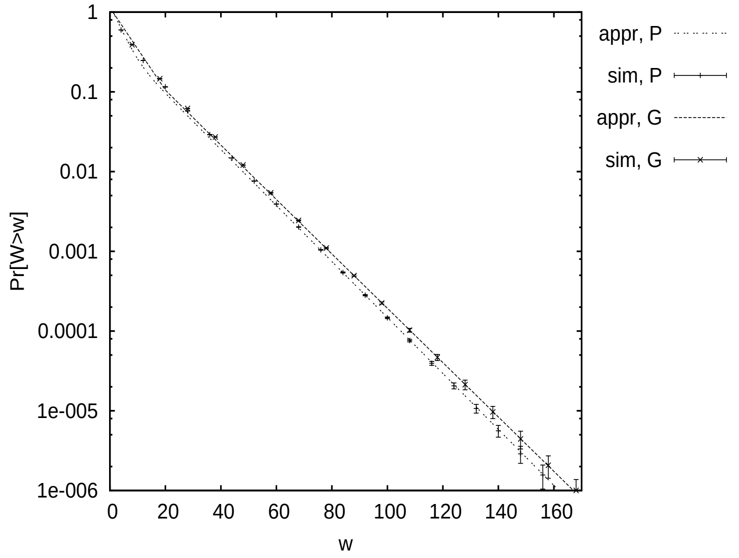


(f) several β's; $c = 10$, $l = 5$, $\rho = 0.3$, Poisson arrivals, geometric services mean $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.35: Evaluation of approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ by comparing it with exact formula (5.46) (3)

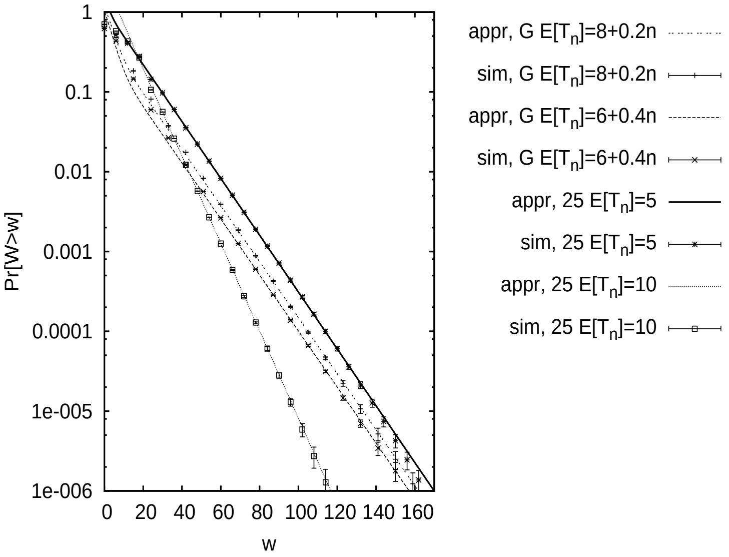


(a) several loads; $c = 10$, $l = 5$, $\beta = 0.05$, Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

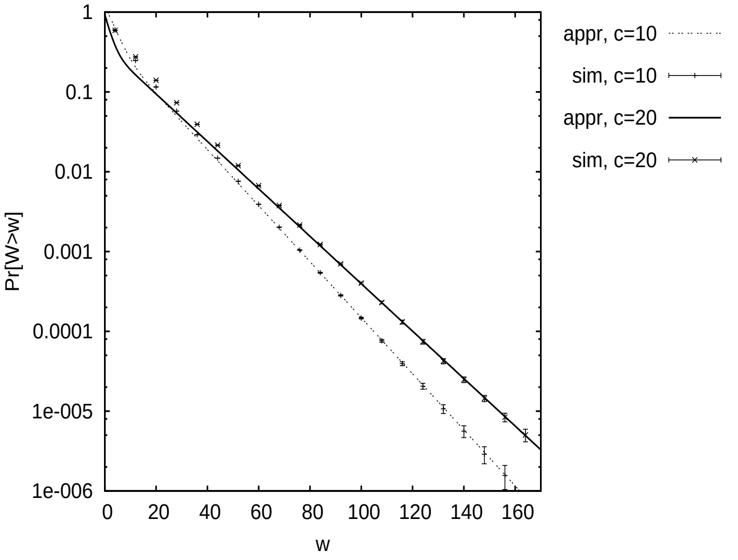


(b) several $A(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.5$, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.36: Evaluation of approximation formula (5.36) for $\Pr[W > w]$; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used (1)

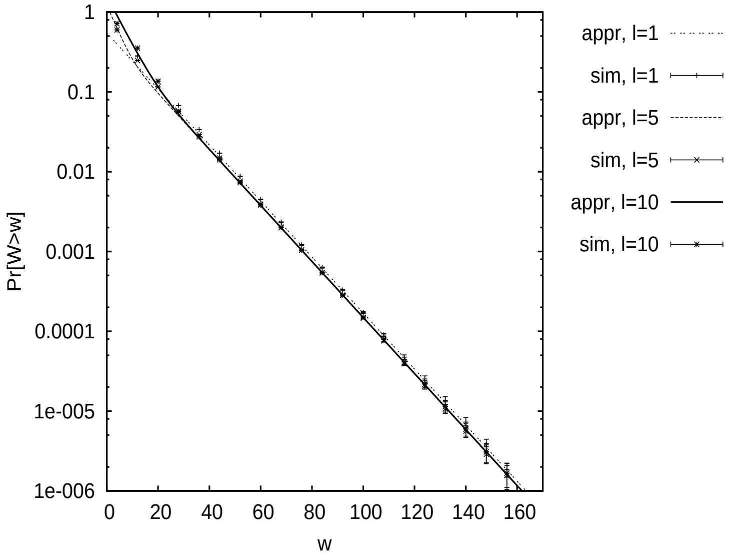


(c) several $T_n(z)$'s; $c = 10$, $l = 5$, $\beta = 0.05$, $\rho = 0.5$, Poisson arrivals

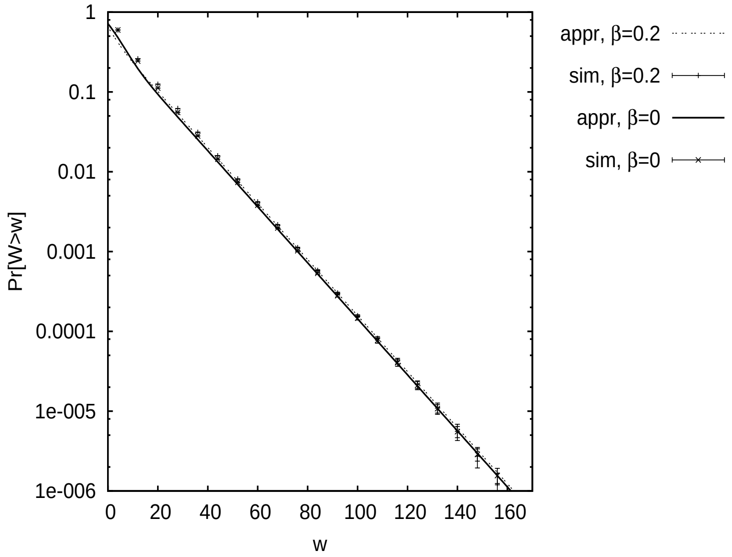


(d) several c's; $l = 5$, $\beta = 0.05$, $\rho = 0.5$, Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.37: Evaluation of approximation formula (5.36) for $\Pr[W > w]$; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used (2)

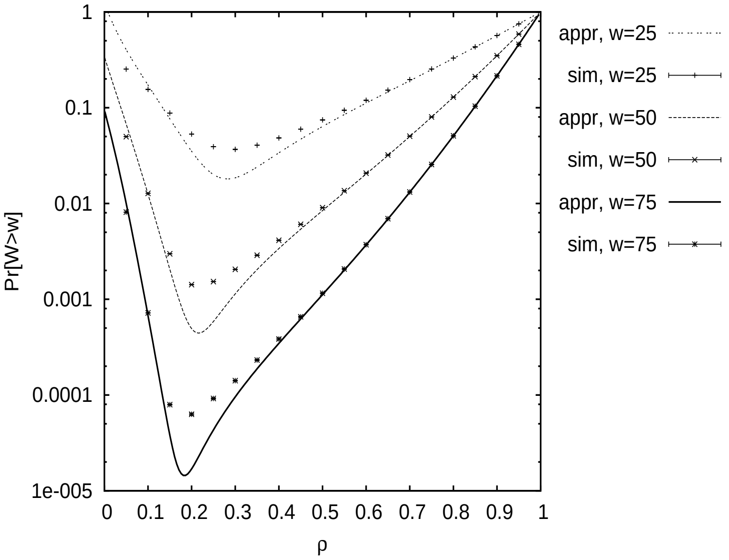


(e) several l's; $c = 10$, $\beta = 0.05$, $\rho = 0.5$, Poisson arrivals, geometric services mean $\mathrm{E}[T_n] = 8 + 0.2n$

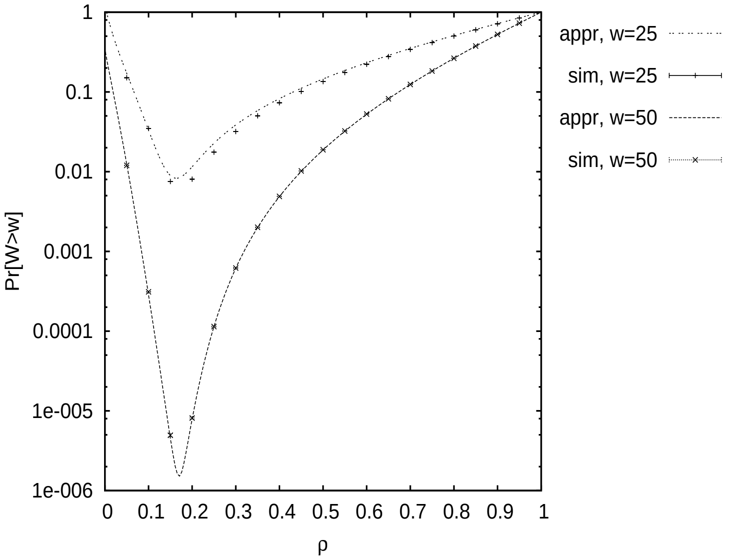


(f) several β's; $c = 10$, $l = 5$, $\rho = 0.5$, Poisson arrivals, geometric services mean $\mathrm{E}[T_n] = 8 + 0.2n$

Figure 5.38: Evaluation of approximation formula (5.36) for $\Pr[W > w]$; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used (3)

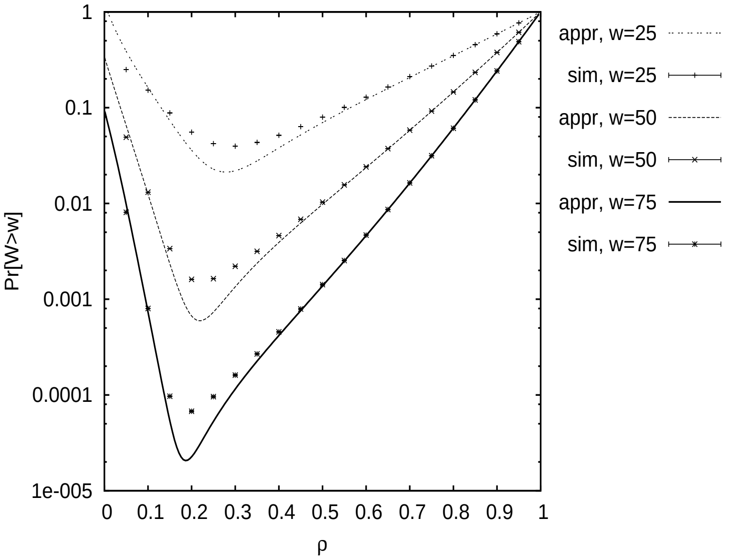


(a) Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$; $c = 10$, $l = 5$, $\beta = 0.05$

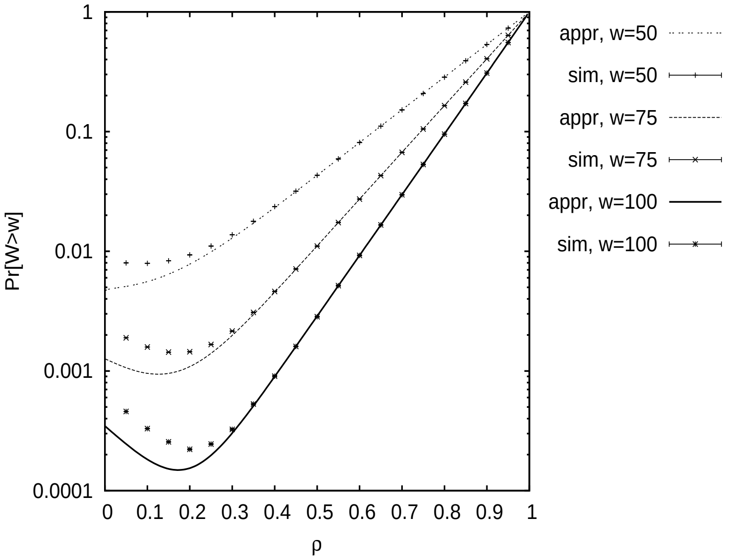


(b) Poisson arrivals, 1 or 25 slots service $\mathrm{E}[T_n] = 5$; $c = 10$, $l = 5$, $\beta = 0.05$

Figure 5.39: Evaluation of approximation formula (5.36) for $\Pr[W > w]$; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used except in (d) where $A^{'}(0) = 0$; $c = 10$, $l = 5$, $\beta = 0.05$ (1)

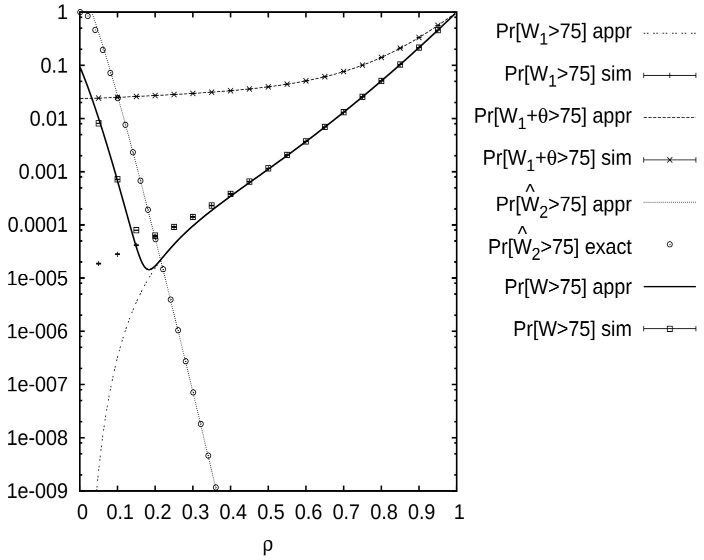


(c) geometric arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$; $c = 10$, $l = 5$, $\beta = 0.05$

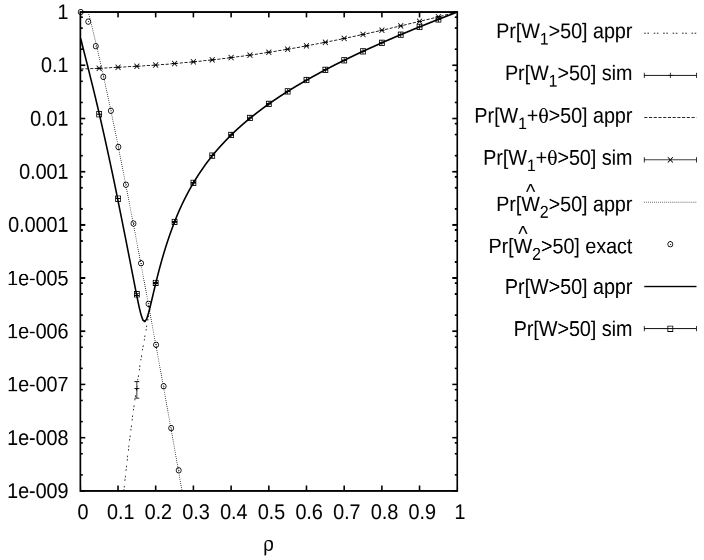


(d) c-centered arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$; $c = 10$, $l = 5$, $\beta = 0.05$

Figure 5.40: Evaluation of approximation formula (5.36) for $\Pr[W > w]$; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used except in (d) where $A^{'}(0) = 0$; $c = 10$, $l = 5$, $\beta = 0.05$ (2)

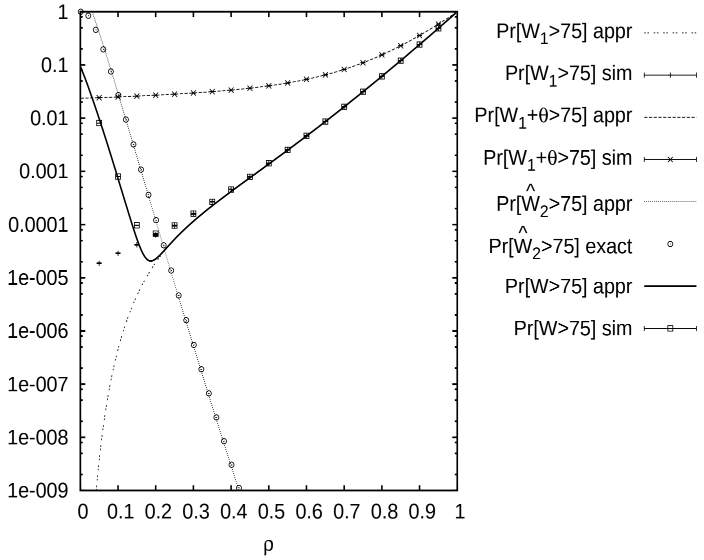


(a) Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$; $c = 10$, $l = 5$, $\beta = 0.05$

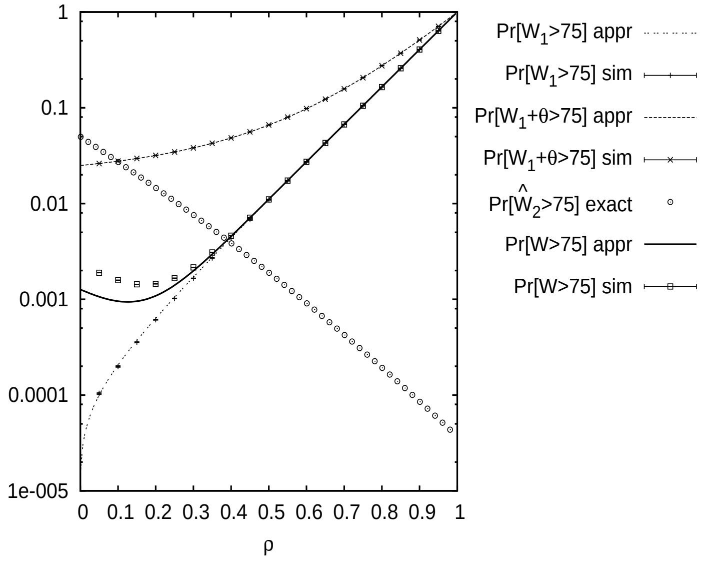


(b) Poisson arrivals, 1 or 25 slots service $\mathrm{E}\left[T_n\right] = 5$; $c = 10$, $l = 5$, $\beta = 0.05$

Figure 5.41: Influence of the load on the approximation formulas; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used except in (d) where $A^{'}(0) = 0$ (1)



(c) geometric arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$; $c = 10$, $l = 5$, $\beta = 0.05$



(d) c-centered arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$; $c = 10$, $l = 5$, $\beta = 0.05$

Figure 5.42: Influence of the load on the approximation formulas; approximation formula (5.47) for $\Pr\left[\hat{W}_2 > w\right]$ is used except in (d) where $A^{'}(0) = 0$ (2)

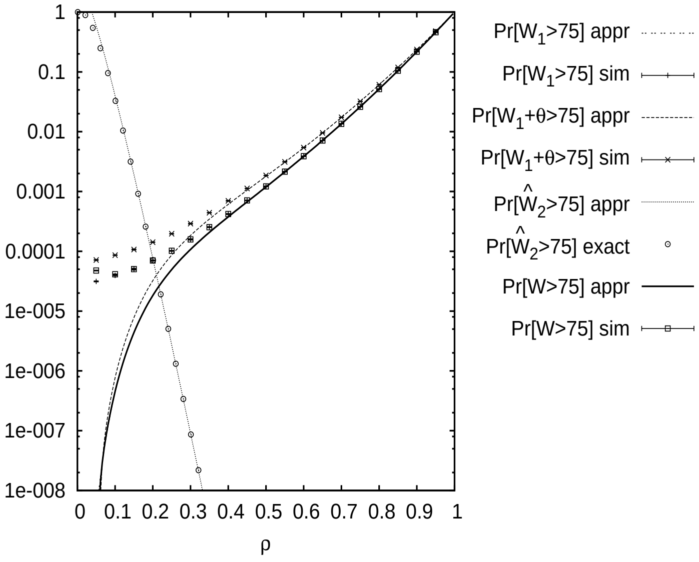


Figure 5.43: Influence of the load on the approximation formulas; Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 8 + 0.2n$, $c = 10$, $l = 5$, $\beta = 0.2$

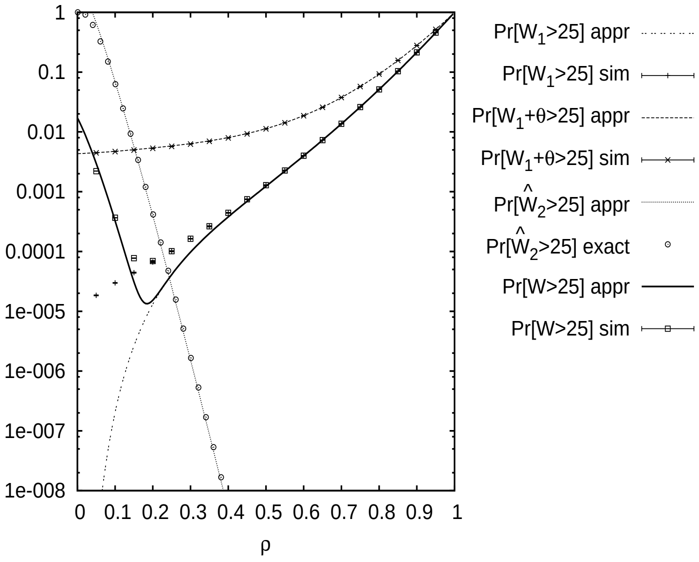


Figure 5.44: Influence of the load on the approximation formulas; Poisson arrivals, geometric services $\mathrm{E}\left[T_n\right] = 3 + 0.1n$, $c = 10$, $l = 5$, $\beta = 0.2$

Chapter 6

Correlated arrivals

6.1 Preface

In the previous chapters, we have studied the buffer content and the customer delay in a versatile batch-service queueing model. However, we have assumed that the number of customer arrivals during consecutive slots was independent and identically distributed (IID), with common PGF $A(z)$, whereas in many real-life circumstances, customer arrivals do not occur independently from each other. For instance, in modern telecommunication systems, a traffic source which is inactive in a given time slot is very likely to remain inactive for a long time (or during a large number of time slots) (see e.g. [60]). The purpose of this chapter is to study the buffer content in the same model as before, except that we also include correlation in the arrival process.

In order to cope with the correlated nature of arrivals, the Markovian arrival process (MAP) can be adopted. In case of a MAP, the probability of having an arrival depends on a background state which is governed by a Markov chain. Several variants of MAP exist: in case of a BMAP, customers arrive in batches instead of individually, whereas D-MAP and D-BMAP represent the discrete-time analogues of MAP and BMAP. Queueing models with MAP (or variants) have been studied extensively in the past, for instance the MAP is considered in [4], [6], [10], [21], [71], [78] and [81], the D-MAP is covered in [31], [35], [67], [111], [112], the BMAP is studied in [1], [11], [56], [85]–[87], [92], [93], [100] and [22], [31], [48], [59], [68], [76], [77], [99], [116] deal with D-BMAP.

Although batch-service queueing models and models with MAP (or variants) have been analyzed separately to a great extent, the combination has attracted much less attention. Exceptions are [12], [34], [64], [66], [104]. Gupta and Laxmu [64] studied the queue content at various epochs in the MAP/$G^{(a,b)}$/1/N queue. Chaudhry and Gupta [34] translated the analysis from [64] to discrete time, resulting in the analysis of the D-MAP/$G^{(a,b)}$/1/N queue. Gupta and Sikdar [66] extended [64] so that single vacations are included and Sikdar and Gupta [104] further extended this research to multiple vacations. Finally, Banik

[12] analyzed the queue content at various epochs in the BMAP/$G^{(a,b)}$/1/N and BMAP/$MSP^{(a,b)}$/1/N systems. Our work differs from these papers in several aspects. First, we consider the D-BMAP, which is more applicable in a telecommunications context due to the discrete nature of the information units that are typically used. Second, we include a dependency between the service time of a batch and the number of items within it. This is closer to reality, since the transmission time of a batch of information packets is typically longer when the batch contains more packets. Also in other application areas, this might be the case. Thirdly, we incorporate a timing mechanism, that avoids excessive delays due to postponing service until the service threshold is reached. This mechanism is of importance when the customers represent for instance real-time data packets. Further, we deduce an additional set of performance quantities compared to [12], [34], [64], [66], [104], where the queue content is established at service completion, pre-arrival and random times. We compute the system content (i.e. the number of customers in the entire system, thus those in service included) at random slot boundaries, the queue content at random slot boundaries, the server content at random slot marks, the system content at service completion times, the number of customers in a served batch, the probability that the server processes a batch during a random slot and the queue content when the server is not processing. The number of customers in a served batch, for instance, is of major concern for practitioners, as it gives a clear indication of the efficiency of the server. Finally, we evaluate more thoroughly the influence of correlation on the behaviour of batch-service queueing systems and more specifically, we investigate the influence on the optimal service threshold.

The work in this chapter is mainly based on our paper [46]. The model considered in [46] is a small extension of the model in this dissertation, in the sense that the probability that service is initiated anyway when not enough customers are present, is dependent on the number of available customers. We omit this dependency in this chapter, in order to be uniform with previous chapters. Paper [46], in turn, is based on our conference paper [45], where we have studied the system content in a batch-service queueing model with a service threshold, with geometrically distributed service times that are independent of the number of served customers, and with a customer arrival process modelled by a D-BMAP. Paper [46] is an extension of [45], in the sense that a more versatile model is considered with service times that are generally distributed and dependent on the number of items in a served batch. In addition, a timing mechanism is included, to avoid excessive delays due to postponing service until the service threshold is reached. Furthermore, a fundamental formula is established from which various quantities related to the number of customers in the queue and server at specific time instants can be deduced, instead of only the system content at random slot boundaries as in [45].

This chapter is structured as follows. The D-BMAP is discussed in section 6.2. Next, in section 6.3, the fundamental formula is established, from which various quantities related to the buffer content are extracted in section 6.4. Further, we discuss how performance measures can be calculated from these quantities in section 6.5 and finally we investigate in section 6.6 the effect of correlation in the arrival process on the behaviour of the system through some numerical examples.

6.2 D-BMAP

Let A_k again represent the number of customer arrivals during slot k. Whereas in previous chapters, we have assumed that the sequence $\{A_k\}_{k\geq 1}$ consists of independent and identically distributed random variables, with common PGF $A(z)$, we now consider correlated (dependent) arrivals. We therefore adopt a D-BMAP (discrete-batch Markovian arrival process), whereby the arrival process is governed by an underlying homogeneous first-order Markov chain, in the sense that the **number of customer arrivals during a slot depends on the transition of the underlying first-order Markov chain**. Let us denote the state of the Markov chain during slot k by τ_k and assume that the Markov chain has a finite number of states N. The arrival process is completely defined by the values $a(n, j|i)$:

$$a(n,j|i) \triangleq \lim_{k\to\infty} \Pr\left[A_k = n, \tau_{k+1} = j|\tau_k = i\right] \ , \qquad n \geq 0; i,j \in \{1,\ldots,N\} \ ,$$

denoting the probability that if the background state is i during a slot, there are n arrivals during this slot and the background state during the next slot is j. We put these probabilities in an $N \times N$ **matrix generating function** $\mathbf{A}(z)$, whose entries are defined as follows:

$$[\mathbf{A}(z)]_{ij} \triangleq \sum_{n=0}^{\infty} a(n,j|i)z^n \ , \qquad i,j \in \{1,\ldots,N\} \ . \tag{6.1}$$

The advantage of working with probability generating matrix (6.1) is twofold: it completely describes the arrival process and it is convenient throughout the analysis. The following information can be extracted from $\mathbf{A}(z)$:

- Transition probabilities p_{ij} of the underlying Markov chain:

  $$p_{ij} \triangleq \lim_{k\to\infty} \Pr\left[\tau_{k+1} = j|\tau_k = i\right] = [\mathbf{A}(1)]_{ij} \ .$$

- Stationary distribution $1 \times N$ vector $\boldsymbol{\pi}$ of the state:

  $$[\boldsymbol{\pi}]_i \triangleq \lim_{k\to\infty} \Pr\left[\tau_k = i\right] \ , \qquad 1 \leq i \leq N \ ,$$

  is the solution of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{A}(1)$ and the normalization condition $\boldsymbol{\pi}\mathbf{1} = 1$, with $\mathbf{1}$ the $N \times 1$ column vector whose N entries are equal to 1.

- Conditional PGF of the number of arrivals given that the background state during that slot equals i:

  $$A_i(z) \triangleq \lim_{k\to\infty} \sum_{n=0}^{\infty} \Pr\left[A_k = n|\tau_k = i\right] z^n = [\mathbf{A}(z)\mathbf{1}]_i \ .$$

- Mean arrival rate λ:

  $$\lambda = \boldsymbol{\pi}\mathbf{A}^{'}(1)\mathbf{1} \ ,$$

  whereby

  $$[\mathbf{A}^{'}(1)]_{ij} = \sum_{n=0}^{\infty} a(n,j|i)n = \left.\frac{\mathrm{d}}{\mathrm{d}z}[\mathbf{A}(z)]_{ij}\right|_{z=1} \ .$$

The matrix generating function $\mathbf{A}(z)$ is convenient to deal with as the matrix generating function of the number of arrivals during two consecutive slots, say $\mathbf{A}_2(z)$, is easily extracted from $\mathbf{A}(z)$ by summing over the "intermediate state" (τ_{k+1}) and taking into account that the number of arrivals during a slot is only dependent on the state of the current and next slot:

$$\begin{aligned}
[\mathbf{A}_2(z)]_{ij} &\triangleq \lim_{k\to\infty} \sum_{n=0}^{\infty} \Pr\left[A_k + A_{k+1} = n, \tau_{k+2} = j|\tau_k = i\right] z^n \\
&= \lim_{k\to\infty} \sum_{n=0}^{\infty} \sum_{m=0}^{n} \sum_{i^{'}=1}^{N} \Pr\left[A_k = m, \tau_{k+1} = i^{'}|\tau_k = i\right] z^m \\
&\quad \Pr\left[A_{k+1} = n-m, \tau_{k+2} = j|\tau_{k+1} = i^{'}\right] z^{n-m} \\
&= \lim_{k\to\infty} \sum_{i^{'}=1}^{N} \sum_{m=0}^{\infty} \Pr\left[A_k = m, \tau_{k+1} = i^{'}|\tau_k = i\right] z^m \\
&\quad \sum_{n=m}^{\infty} \Pr\left[A_{k+1} = n-m, \tau_{k+2} = j|\tau_{k+1} = i^{'}\right] z^{n-m} \\
&= \sum_{i^{'}=1}^{N} [\mathbf{A}(z)]_{ii^{'}} [\mathbf{A}(z)]_{i^{'}j} \ ,
\end{aligned}$$

which is the entry at the i-th row and j-th column of the matrix product $\mathbf{A}(z)^2$. Consequently, as

$$[\mathbf{A}_2(z)]_{ij} = [\mathbf{A}(z)^2]_{ij} \ ,$$

for all $i, j \in \{1, \ldots, N\}$, it holds that $\mathbf{A}_2(z) = \mathbf{A}(z)^2$. Analogously, one can prove by mathematical induction that the matrix generating function of the number of arrivals during n consecutive slots equals $\mathbf{A}(z)^n$ and consequently, that the number of arrivals during the service of n customers equals $T_n(\mathbf{A}(z)) \triangleq \sum_{k=0}^{\infty} \Pr\left[T_n = k\right] \mathbf{A}(z)^k$.

The radius of convergence of $\mathbf{A}(z)$ is designated by $\Re_{\mathbf{A}}$ and is equal to

$$\Re_{\mathbf{A}} \triangleq \min_{i,j} \Re_{\mathbf{A}_{ij}} \ ,$$

with $\Re_{\mathbf{A}_{ij}}$ the radius of convergence of $[\mathbf{A}(z)]_{ij}$. Hence, each of the entries of $\mathbf{A}(z)$ is an analytic function in the open disk $\{z \in \mathbb{C} : |z| < \Re_{\mathbf{A}}\}$.

During the analysis in the subsequent sections, we will make use of **spectral decomposition**[1]. We thereby assume that $\mathbf{A}(z)$ is diagonalizable, i.e.

[1] For a good introduction on matrix algebra and more specifically spectral decomposition we strongly recommend the book [89].

$\mathbf{A}(z)$ can be factorized as

$$\mathbf{A}(z) = \mathbf{R}(z)\mathbf{\Lambda}(z)\mathbf{R}^{-1}(z) \ , \tag{6.2}$$

with $\mathbf{\Lambda}(z)$ a diagonal matrix. It can be proved (see e.g. [89]) that $\mathbf{A}(z)$ is diagonalizable if and only if it possesses a complete set of eigenvectors, that the columns $\mathbf{r}_j(z)$ of $\mathbf{R}(z)$ then constitute a complete set of right eigenvectors and that the diagonal entries $\lambda_i(z)$ of $\mathbf{\Lambda}(z)$ are the eigenvalues of $\mathbf{A}(z)$, so that each $(\lambda_j(z), \mathbf{r}_j(z))$ is an eigenpair for $\mathbf{A}(z)$:

$$\mathbf{A}(z)\mathbf{r}_j(z) = \lambda_j(z)\mathbf{r}_j(z) \ , \qquad 1 \leq j \leq N \ .$$

Note that the eigenvectors are unique upon some factor, which we can, without loss of generality, fix by making the convention that the row sums of either $\mathbf{R}(z)$ or $\mathbf{R}^{-1}(z)$ are equal to one (the former implies the latter and vice versa):

$$\mathbf{R}(z)\mathbf{1} = \mathbf{1} \Leftrightarrow \mathbf{R}^{-1}(z)\mathbf{1} = \mathbf{1} \ .$$

This convention will turn out to be convenient throughout the calculation of performance measures. Next, relation (6.2) implies that

$$\mathbf{A}(z)^n = \mathbf{R}(z)\mathbf{\Lambda}(z)^n\mathbf{R}^{-1}(z) \ ,$$

$$T_c(\mathbf{A}(z)) = \mathbf{R}(z)T_c(\mathbf{\Lambda}(z))\mathbf{R}^{-1}(z) \ ,$$

which means that

$$\mathbf{A}(z)^n\mathbf{r}_j(z) = \lambda_j(z)^n\mathbf{r}_j(z) \ , \qquad 1 \leq j \leq N \ ,$$

and

$$T_c(\mathbf{A}(z))\mathbf{r}_j(z) = T_c(\lambda_j(z))\mathbf{r}_j(z) \ , \qquad 1 \leq j \leq N \ .$$

In other words, each $\mathbf{r}_j(z)$ is a right eigenvector of $\mathbf{A}(z)^n$ and $T_c(\mathbf{A}(z))$ as well, with corresponding eigenvalues $\lambda_j(z)^n$ and $T_c(\lambda_j(z))$ respectively.

Next, since $\mathbf{A}(z)$ is a matrix with positive entries for all $z \in]0, \Re_{\mathbf{A}}[$, it has one real and positive eigenvalue that exceeds the moduli of all other eigenvalues for these values of z ([89]). This eigenvalue is called the **Perron-Frobenius (PF) eigenvalue** and we let $\lambda_1(z)$ represent that eigenvalue. The PF eigenvalue and its corresponding right eigenvector satisfy $\lambda_1(1) = 1$, $\mathbf{r}_1(1) = \mathbf{1}$. In addition, it can be proved that $\lambda_1^{'}(1) = \lambda$. Indeed, it holds that $\mathbf{A}(z)\mathbf{r}_1(z) = \lambda_1(z)\mathbf{r}_1(z)$. Taking the first derivative at $z = 1$ and invoking $\lambda_1(1) = 1$ and $\mathbf{r}_1(1) = \mathbf{1}$ yields

$$\mathbf{A}^{'}(1)\mathbf{1} + \mathbf{A}(1)\mathbf{r}_1^{'}(1) = \lambda_1^{'}(1)\mathbf{1} + \mathbf{r}_1^{'}(1) \ . \tag{6.3}$$

Multiplying (6.3) to the left with $\boldsymbol{\pi}$, the steady-state vector of the state of the underlying Markov chain, relying on $\boldsymbol{\pi}\mathbf{A}(1) = \boldsymbol{\pi}$, $\boldsymbol{\pi}\mathbf{A}^{'}(1)\mathbf{1} = \lambda$ and $\boldsymbol{\pi}\mathbf{1} = 1$, produces

$$\lambda_1^{'}(1) = \lambda \ .$$

Before closing this section, we define the **vector generating function** $\mathbf{X}(z)$ of a random variable X that depends on the background state (X_k represents the value of X at slot mark k) as the $1 \times N$ vector whose entries are defined as follows:

$$\begin{aligned}[\mathbf{X}(z)]_j &\triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{X_k}\{\tau_k = j\}\right] \\ &= \lim_{k\to\infty} \mathrm{E}\left[z^{X_k}|\tau_k = j\right] \Pr\left[\tau_k = j\right] \ .\end{aligned}$$

Note that the PGF $X(z)$ of X can easily be extracted from the vector generating function $\mathbf{X}(z)$: $X(z) = \mathbf{X}(z)\mathbf{1}$, and that $\mathbf{X}(1)$ is equal to the steady-state vector $\boldsymbol{\pi}$ of the background state.

6.3 Joint vector generating function

In this section, we compute the $1 \times N$ steady state joint vector generating function $\mathbf{V}(z, x, y)$ of the queue content, the server content and the remaining service time:

$$[\mathbf{V}(z,x,y)]_j \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k} x^{S_k} y^{R_k}\{\tau_k = j\}\right] \ ,$$

with Q_k (S_k) the queue (server) content and R_k the remaining service time at slot boundary k. We commence by writing down the system equations, which express the relation between $(Q_{k+1}, S_{k+1}, R_{k+1})$ and (Q_k, S_k, R_k):

$$(Q_{k+1}, S_{k+1}, R_{k+1}) =$$
$$\begin{cases} (Q_k + A_k, S_k, R_k - 1) & \text{if } R_k > 1 \\ (0, Q_k + A_k, T_{Q_k+A_k}) & \text{if } R_k \le 1 \text{ and } l \le Q_k + A_k < c \\ (Q_k + A_k - c, c, T_c) & \text{if } R_k \le 1 \text{ and } Q_k + A_k \ge c \\ (0, Q_k + A_k, T_{Q_k+A_k}) & \begin{array}{l}\text{if } R_k \le 1,\ Q_k + A_k < l \text{ and service starts} \\ \text{(with probability } \beta)\end{array} \\ (Q_k + A_k, 0, 0) & \begin{array}{l}\text{if } R_k \le 1,\ Q_k + A_k < l \text{ and service does} \\ \text{not start (with probability } 1-\beta)\end{array} \end{cases}$$

Indeed, in the first case, the service continues during slot $k+1$, so that customers that have arrived during slot k are stored in the queue. In the other cases, the server is available at slot mark $k+1$. Whether a new service is initiated or not is described by the service policy, which is described in section 1.5, and is thus dependent on the number of available customers.
The system equations can be translated into vector generating functions as follows:

$$\begin{aligned}[\mathbf{V}_{k+1}(z,x,y)]_j \triangleq & \mathrm{E}\left[z^{Q_{k+1}} x^{S_{k+1}} y^{R_{k+1}}\{\tau_{k+1} = j\}\right] \\ = & \frac{1}{y}\mathrm{E}\left[z^{Q_k+A_k} x^{S_k} y^{R_k}\{R_k > 1, \tau_{k+1} = j\}\right] \\ & + \mathrm{E}\left[x^{Q_k+A_k} y^{T_{Q_k+A_k}}\{R_k \le 1, l \le Q_k + A_k < c, \tau_{k+1} = j\}\right] \\ & + \left(\frac{x}{z}\right)^c T_c(y)\mathrm{E}\left[z^{Q_k+A_k}\{R_k \le 1, Q_k + A_k \ge c, \tau_{k+1} = j\}\right] \\ & + \beta\mathrm{E}\left[x^{Q_k+A_k} y^{T_{Q_k+A_k}}\{R_k \le 1, Q_k + A_k < l, \tau_{k+1} = j\}\right] \\ & + (1-\beta)\mathrm{E}\left[z^{Q_k+A_k}\{R_k \le 1, Q_k + A_k < l, \tau_{k+1} = j\}\right] \ . \end{aligned} \tag{6.4}$$

We now calculate each term from the right-hand-side of (6.4) separately. We therefore introduce the $1 \times N$ row vectors $\mathbf{q}_{0k}(n)$, $\mathbf{d}_k(n)$ and $\mathbf{F}_k(z, x)$ as

$$[\mathbf{q}_{0k}(n)]_j \triangleq \Pr\left[Q_k = n, R_k = 0, \tau_k = j\right] \ , \tag{6.5}$$

$$[\mathbf{d}_k(n)]_j \triangleq \Pr\left[Q_k + A_k = n, R_k \le 1, \tau_{k+1} = j\right] \ , \tag{6.6}$$

$$[\mathbf{F}_k(z,x))]_j \triangleq \mathrm{E}\left[z^{Q_k} x^{S_k}\{R_k = 1, \tau_k = j\}\right] \ . \tag{6.7}$$

Let us start with the first term from (6.4). We take the sum over all possible states τ_k during slot k:

$$\mathrm{E}\left[z^{Q_k+A_k} x^{S_k} y^{R_k}\{R_k > 1, \tau_{k+1} = j\}\right] = \sum_{i=1}^{N} \mathrm{E}\left[z^{Q_k+A_k} x^{S_k} y^{R_k}\{R_k > 1, \tau_{k+1} = j, \tau_k = i\}\right] \ .$$

As A_k is independent of Q_k, S_k and R_k when τ_k and τ_{k+1} are given, this expresssion transforms into

$$\begin{aligned}\mathrm{E}\left[z^{Q_k+A_k}x^{S_k}y^{R_k}\{R_k>1,\tau_{k+1}=j\}\right] = &\sum_{i=1}^{N}\mathrm{E}\left[z^{Q_k}x^{S_k}y^{R_k}\{R_k>1\}|\tau_{k+1}=j,\tau_k=i\right]\\ &.\ \Pr\left[\tau_k=i\right]\mathrm{E}\left[z^{A_k}|\tau_{k+1}=j,\tau_k=i\right]\\ &.\ \Pr\left[\tau_{k+1}=j|\tau_k=i\right]\ .\end{aligned}$$

Note that Q_k, S_k and R_k are independent of τ_{k+1} if τ_k is given. Indeed, Q_k, S_k and R_k are influenced by A_{k-1}, which is not dependent of τ_{k+1} if τ_k is known. As a result, we find, by invoking the definitions of $\mathbf{A}(z)$, $\mathbf{V}_k(z,x,y)$ and (6.5)-(6.7) and by taking into account that the underlying Markov chain is time homogeneous:

$$\begin{aligned}&\mathrm{E}\left[z^{Q_k+A_k}x^{S_k}y^{R_k}\{R_k>1,\tau_{k+1}=j\}\right] =\\ &\qquad\sum_{i=1}^{N}\left[\mathbf{V}_k(z,x,y)-\sum_{n=0}^{l-1}\mathbf{q}_{0k}(n)z^n - y\mathbf{F}_k(z,x)\right]_i [\mathbf{A}(z)]_{ij}\ ,\end{aligned}$$

which is nothing else than a matrix multiplication. Hence,

$$\begin{aligned}&\mathrm{E}\left[z^{Q_k+A_k}x^{S_k}y^{R_k}\{R_k>1,\tau_{k+1}=j\}\right] =\\ &\qquad\left[\left\{\mathbf{V}_k(z,x,y)-\sum_{n=0}^{l-1}\mathbf{q}_{0k}(n)z^n - y\mathbf{F}_k(z,x)\right\}\mathbf{A}(z)\right]_j\ .\end{aligned} \tag{6.8}$$

The third term can be established analogously as the first, which yields:

$$\begin{aligned}&\mathrm{E}\left[z^{Q_k+A_k}\{R_k\leq 1, Q_k+A_k\geq c,\tau_{k+1}=j\}\right] =\\ &\qquad\left[\mathbf{F}_k(z,1)\mathbf{A}(z)+\sum_{n=0}^{l-1}\mathbf{q}_{0k}(n)z^n\mathbf{A}(z)-\sum_{n=0}^{c-1}\mathbf{d}_k(n)z^n\right]_j\ .\end{aligned} \tag{6.9}$$

The other terms are easier to calculate, because we just have to rely on definition (6.6) of $\mathbf{d}_k(n)$. As a result, we find for respectively the second, fourth and fifth term:

$$\mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\{R_k\leq 1, l\leq Q_k+A_k<c,\tau_{k+1}=j\}\right] = \left[\sum_{n=l}^{c-1}\mathbf{d}_k(n)x^nT_n(y)\right]_j\ , \tag{6.10}$$

$$\mathrm{E}\left[x^{Q_k+A_k}y^{T_{Q_k+A_k}}\{R_k\leq 1, Q_k+A_k<l,\tau_{k+1}=j\}\right] = \left[\sum_{n=0}^{l-1}\mathbf{d}_k(n)x^nT_n(y)\right]_j\ , \tag{6.11}$$

$$\mathrm{E}\left[z^{Q_k+A_k}\{R_k\leq 1, Q_k+A_k<l,\tau_{k+1}=j\}\right] = \left[\sum_{n=0}^{l-1}\mathbf{d}_k(n)z^n\right]_j\ . \tag{6.12}$$

The substitution of (6.8)-(6.12) in (6.4) produces in the steady state

$$\begin{aligned}\mathbf{V}(z,x,y) =& \frac{1}{y}\left\{\mathbf{V}(z,x,y) - \sum_{n=0}^{l-1}\mathbf{q}_0(n)z^n - y\mathbf{F}(z,x)\right\}\mathbf{A}(z)\\ &+ \sum_{n=l}^{c-1}\mathbf{d}(n)x^nT_n(y)\\ &+ \left(\frac{x}{z}\right)^c T_c(y)\left[\mathbf{F}(z,1)\mathbf{A}(z) + \sum_{n=0}^{l-1}\mathbf{q}_0(n)z^n\mathbf{A}(z) - \sum_{n=0}^{c-1}\mathbf{d}(n)z^n\right]\\ &+ \beta\sum_{n=0}^{l-1}\mathbf{d}(n)x^nT_n(y) + (1-\beta)\sum_{n=0}^{l-1}\mathbf{d}(n)z^n \ ,\end{aligned} \tag{6.13}$$

whereby the $1 \times N$ row vectors $\mathbf{q_0}(n)$, $\mathbf{d}(n)$ and $\mathbf{F}(z,x)$ represent the steady-state equivalents of $\mathbf{q}_{0k}(n)$, $\mathbf{d}_k(n)$ and $\mathbf{F}_k(z,x)$:

$$[\mathbf{q_0}(n)]_j \triangleq \lim_{k\to\infty} \Pr\left[Q_k = n, R_k = 0, \tau_k = j\right] \ , \tag{6.14}$$

$$[\mathbf{d}(n)]_j \triangleq \lim_{k\to\infty} \Pr\left[Q_k + A_k = n, R_k \le 1, \tau_{k+1} = j\right] \ , \tag{6.15}$$

$$[\mathbf{F}(z,x))]_j \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{Q_k}x^{S_k}\{R_k = 1, \tau_k = j\}\right] \ .$$

Note that equation (6.13) is similar to expression (2.6) in the independent case, except that PGFs are substituted by matrix or vector generating functions. However, it is important to note that one has to be careful with multiplication now, as matrix and vector multiplications are not commutative. Next, notice that definitions (6.14) and (6.15) imply that

$$\mathbf{q}_0(n) = \mathbf{d}(n)(1-\beta) \ , \qquad 0 \le n \le l-1 \ . \tag{6.16}$$

Substitution of (6.16) in (6.13) produces

$$\begin{aligned}\mathbf{V}(z,x,y)\left[\mathbf{I} - \frac{1}{y}\mathbf{A}(z)\right] =& (1-\beta)\sum_{n=0}^{l-1}\mathbf{d}(n)z^n\left[\mathbf{I} - \frac{\mathbf{A}(z)}{y}\right]\\ &+ \left(\frac{x}{z}\right)^c T_c(y)\sum_{n=0}^{l-1}\mathbf{d}(n)z^n[\mathbf{A}(z) - \mathbf{I}]\\ &+ \beta\sum_{n=0}^{l-1}\mathbf{d}(n)\left[x^nT_n(y)\mathbf{I} - z^n\left(\frac{x}{z}\right)^c T_c(y)\mathbf{A}(z)\right]\\ &+ \left(\frac{x}{z}\right)^c T_c(y)\mathbf{F}(z,1)\mathbf{A}(z) - \mathbf{F}(z,x)\mathbf{A}(z)\\ &+ \sum_{n=l}^{c-1}\mathbf{d}(n)\left[x^nT_n(y) - z^n\left(\frac{x}{z}\right)^c T_c(y)\right] \ ,\end{aligned} \tag{6.17}$$

with $\mathbf{I}$ the $N \times N$ identity matrix. For the purpose of extracting performance measures in the next sections, it turns out to be more convenient to multiply expression (6.17) to the right with $\mathbf{r}_i(z)$, the i-th right eigenvector of $\mathbf{A}(z)$. We obtain

$$\begin{aligned}\left[1-\frac{\lambda_i(z)}{y}\right]\mathbf{V}(z,x,y)\mathbf{r}_i(z) =&(1-\beta)\left[1-\frac{\lambda_i(z)}{y}\right]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n\\ &+\left(\frac{x}{z}\right)^c T_c(y)[\lambda_i(z)-1]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n\\ &+\beta\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[x^nT_n(y)-z^n\left(\frac{x}{z}\right)^cT_c(y)\lambda_i(z)\right]\\ &+\left(\frac{x}{z}\right)^c T_c(y)\lambda_i(z)\mathbf{F}(z,1)\mathbf{r}_i(z)-\lambda_i(z)\mathbf{F}(z,x)\mathbf{r}_i(z)\\ &+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[x^nT_n(y)-z^n\left(\frac{x}{z}\right)^cT_c(y)\right]\quad . \qquad (6.18)\end{aligned}$$

Substituting y by $\lambda_i(z)$ and letting $x\rightarrow 1$ leads to

$$\begin{aligned}\lambda_i(z)\left[z^c-T_c(\lambda_i(z))\right]\mathbf{F}(z,1)\mathbf{r}_i(z) =&T_c(\lambda_i(z))[\lambda_i(z)-1]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n\\ &+\beta\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\lambda_i(z)\right]\\ &+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\right]\quad . \qquad (6.19)\end{aligned}$$

Substituting y by $\lambda_i(z)$ in (6.18) and appealing to (6.19) yields

$$\begin{aligned}&z^c\lambda_i(z)\left[z^c-T_c(\lambda_i(z))\right]\mathbf{F}(z,x)\mathbf{r}_i(z)\\ &\quad=z^cx^cT_c(\lambda_i(z))[\lambda_i(z)-1]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n\\ &\quad+\beta x^cT_c(\lambda_i(z))\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\lambda_i(z)\right]\\ &\quad+x^cT_c(\lambda_i(z))\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cT_n(\lambda_i(z))-z^nT_c(\lambda_i(z))\right]\\ &\quad+\beta[z^c-T_c(\lambda_i(z))]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cx^nT_n(\lambda_i(z))-x^cz^nT_c(\lambda_i(z))\lambda_i(z)\right]\\ &\quad+[z^c-T_c(\lambda_i(z))]\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\left[z^cx^nT_n(\lambda_i(z))-x^cz^nT_c(\lambda_i(z))\right]\quad . \qquad (6.20)\end{aligned}$$

Expressions (6.18)-(6.20) provide enough information to deduce a spectrum of quantities related to the buffer content, which constitutes the subject of the next section. However, formulas (6.18)-(6.20) still contain the unknown vectors $\mathbf{d}(n)$. In order to explain how these can be calculated, set $y=1$ and $x=z$ in (6.18) and rely on (6.19)-(6.20), leading to

$$\begin{aligned}[1-\lambda_i(z)]\left[z^c-T_c(\lambda_i(z))\right]\mathbf{V}(z,z,1)\mathbf{r}_i(z) =&(z^c-1)T_c(\lambda_i(z))[1-\lambda_i(z)]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n\\ &+\beta\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{g}_{n,i}(z)+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{h}_{n,i}(z)\quad , \qquad (6.21)\end{aligned}$$

with

$$\begin{aligned}\mathbf{g}_{n,i}(z) \triangleq & [(z^n - z^c)T_n(\lambda_i(z))T_c(\lambda_i(z)) + z^n(z^c-1)T_c(\lambda_i(z))\lambda_i(z) \\ & -z^c(z^n-1)T_n(\lambda_i(z))]\,\mathbf{r}_i(z) \ ,\end{aligned} \tag{6.22}$$

$$\begin{aligned}\mathbf{h}_{n,i}(z) \triangleq & [T_n(\lambda_i(z))z^c\{(1-z^n) - T_c(\lambda_i(z))\} \\ & -T_c(\lambda_i(z))z^n\{(1-z^c) - T_n(\lambda_i(z))\}]\,\mathbf{r}_i(z) \ .\end{aligned} \tag{6.23}$$

Unlike the case of independent arrivals, it is impossible to construct an irrefutable mathematical proof, based on Rouché's theorem, to show that each of the equations $z^c - T_c(\lambda_i(z)) = 0 \ , i = 1, \ldots, N$, necessarily has c solutions inside the closed complex unit disk. Nevertheless, an example where this is not the case has not been encountered up to now, and, to the best of our knowledge, such an example, if it exists, has yet to be constructed. Hence, for practical purposes, we can venture to state that the above equation has indeed c solutions inside the closed complex unit disk for each value of i, provided that the equilibrium condition $\rho < 1$ holds.
Let us characterise the k-th solution of the i-th equation by $z_{i,k}$. As $\lambda_1(1) = 1$, one of the zeros of $z^c - T_c(\lambda_1(z))$ equals one. Without loss of generality, we let $z_{1,1}$ be that zero. As $\mathbf{V}(z, z, 1)$ is analytic inside the closed complex unit disk, the right-hand-side of (6.21) must also vanish at these zeroes. This observation leads to $Nc - 1$ linear equations in the $1 \times N$ vectors $\mathbf{d}(n)$, $n = 0, \ldots, c-1$. The zero $z_{1,1}$ cannot be used as it produces the trivial equation $0 = 0$. Fortunately, we can resort to the normalization condition to obtain another equation. Derivating (6.21) twice at $z = 1$ for $i = 1$ and taking into account that $\mathbf{r}_1(1) = \mathbf{1}$, $\mathbf{V}(z, z, 1)\mathbf{1} = 1$ and $\lambda_1'(1) = \lambda$, produces the normalization condition

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}z^2} f_1(z)\right|_{z=1} = -2\lambda c(1-\rho) \ ,$$

with $f_i(z)$ the right-hand-side of (6.21).

6.4 Quantities related to the buffer content

6.4.1 System content at random slot boundaries

As the system content U equals the sum of the queue and the server content, its vector generating function $\mathbf{U}(z)$ is equal to $\mathbf{V}(z, z, 1)$. Hence, substituting $\mathbf{V}(z, z, 1)$ by $\mathbf{U}(z)$ in expression (6.21) yields

$$\begin{aligned}[1-\lambda_i(z)]\,[z^c - T_c(\lambda_i(z))]\,\mathbf{U}(z)\mathbf{r}_i(z) = & (z^c-1)T_c(\lambda_i(z))[1-\lambda_i(z)]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\ & + \beta\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{g}_{n,i}(z) + \sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{h}_{n,i}(z) \ ,\end{aligned} \tag{6.24}$$

whereby $\mathbf{g}_{n,i}(z)$ and $\mathbf{h}_{n,i}(z)$ are defined by respectively (6.22) and (6.23).

6.4.2 Queue content at random slot boundaries

The vector generating function $\mathbf{Q}(z)$ of the queue content at random slot boundaries is found by summing out both the server content and the remaining service time. Hence, letting $y \to 1$ and $x \to 1$ in (6.18) and applying (6.19), we find

$$\begin{aligned}
&[1-\lambda_i(z)]\,[z^c - T_c(\lambda_i(z))]\,\mathbf{Q}(z)\mathbf{r}_i(z) \\
&\quad=(z^c-1)[1-\lambda_i(z)]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)z^n \\
&\quad\quad+\beta\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{r}_i(z)\,[(1-z^n)\{z^c-T_c(\lambda_i(z))\}+(z^c-1)\{z^n\lambda_i(z)-T_n(\lambda_i(z))\}] \\
&\quad\quad+\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{r}_i(z)\,[z^c-z^n+(z^n-1)T_c(\lambda_i(z))+(1-z^c)T_n(\lambda_i(z))]\ . \qquad (6.25)
\end{aligned}$$

6.4.3 System content at service completion times

The system content $\tilde{U}$ at service completion times equals the sum of the queue content at the beginning of the last slot of the service and the customers that have arrived during that slot. Hence, by definition, we get

$$\tilde{\mathbf{U}}(z)\mathbf{r}_i(z) = \lambda_i(z)\frac{\mathbf{F}(z,1)\mathbf{r}_i(z)}{\mathbf{F}(1,1)\mathbf{1}}\ . \qquad (6.26)$$

Note that the following formula is also valid:

$$\tilde{U}(z) = \frac{\mathbf{F}(z,1)\mathbf{A}(z)\mathbf{1}}{\mathbf{F}(1,1)\mathbf{1}}\ .$$

We however prefer formula (6.26) as we have deduced an expression for $\mathbf{F}(z,x)$ $\mathbf{r}_i(z)$ and not for $\mathbf{F}(z,x)$ separately.

6.4.4 Server content at random slot boundaries

The probability generating function $S(z)$ of the server content at random slot boundaries is found by first substituting i by 1 and y by 1 in (6.18), then invoking (6.20) and finally letting $z \to 1$ and thereby applying l'Hôpital's rule twice, eventually resulting in:

$$\begin{aligned}
&S(z)\,[c-\mathrm{E}\,[T_c]\,\lambda] \\
&\quad=(1-\beta)\,[c-\mathrm{E}\,[T_c]\,\lambda]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{1}+\beta\,[c-\mathrm{E}\,[T_c]\,\lambda]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{1}z^n\mathrm{E}\,[T_n] \\
&\quad\quad+[c-\mathrm{E}\,[T_c]\,\lambda]\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{1}z^n\mathrm{E}\,[T_n]+z^c\lambda\mathrm{E}\,[T_c]\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{1} \\
&\quad\quad+z^c\mathrm{E}\,[T_c]\,\beta\sum_{n=0}^{l-1}\mathbf{d}(n)\mathbf{1}\{\mathrm{E}\,[T_n]\,\lambda-n-\lambda\}+z^c\mathrm{E}\,[T_c]\sum_{n=l}^{c-1}\mathbf{d}(n)\mathbf{1}\{\mathrm{E}\,[T_n]\,\lambda-n\}\ . \qquad (6.27)
\end{aligned}$$

6.4.5 Number of customers in a served batch

The number of customers in a random served batch, $\tilde{S}$, is equally distributed as the server content at the last slot of a random service period, which yields

$$\tilde{\mathbf{S}}(z)\mathbf{r}_i(z) = \frac{\mathbf{F}(1,z)\mathbf{r}_i(z)}{\mathbf{F}(1,1)\mathbf{1}}\ . \qquad (6.28)$$

6.4.6 Probability that the server processes

The probability that the server processes during a random slot ensues almost immediately from the definition of $\mathbf{q}_0(n)$:

$$\Pr[\text{server processes}] = 1 - \sum_{n=0}^{l-1} \mathbf{q}_0(n)\mathbf{1} \ . \tag{6.29}$$

6.4.7 Queue content when the server not processes

The vector generating function $\tilde{\mathbf{Q}}(z)$ of the queue content when the server not processes (because none or not enough customers are present to start a new service), is found by taking into account that the server is not processing if and only if the remaining service time equals 0. Hence

$$\tilde{Q}(z) = \frac{\sum_{n=0}^{l-1} \mathbf{q}_0(n)\mathbf{1}z^n}{\mathbf{V}(1,0,0)\mathbf{1}} = \frac{\sum_{n=0}^{l-1} \mathbf{q}_0(n)\mathbf{1}z^n}{\mathbf{V}(1,0,0)\mathbf{r}_1(1)} \ . \tag{6.30}$$

6.5 Performance measures

In section 6.4, we have deduced various quantities related to the buffer content ((6.24)-(6.30)). These formulas allow us to calculate performance measures such as moments and tail probabilities. As compared to the case of independent arrivals, this matter is more complicated now and we therefore briefly explain how the mean value of the system content and the tail probabilities of the queue content can be calculated.

The mean value of the system content is found by differentiating (6.24) three times at $z = 1$ for $i = 1$ and taking into account that $\mathbf{r}_1(1) = \mathbf{1}$, $\lambda_1^{'}(1) = \lambda$ and $\mathbf{U}^{'}(1)\mathbf{1} = \mathrm{E}[U]$, leading to

$$\mathrm{E}[U] = \frac{1}{6\lambda c(\rho-1)} \left[\left. \frac{\mathrm{d}^3}{\mathrm{d}z^3} f_1(z) \right|_{z=1} + 6\lambda c(1-\rho)\mathbf{U}(1)\mathbf{r}_1'(1) \right.$$
$$\left. - 3\lambda^3 T_c''(1) - 6\lambda_1''(1)T_c'(1)\lambda + 3\lambda_1''(1)c + 3\lambda c^2 - 3\lambda c \right] ,$$

with $f_i(z)$ the right-hand-side of (6.24).

Throughout this dissertation, we assume an infinite buffer capacity. Nevertheless, buffers have a finite capacity, which causes customers to get rejected if they arrive when the buffer is full. As a result, the loss ratio - defined as the fraction of customers that are rejected - is an important performance measure. In addition, the buffer capacity, say b, is typically large, and the tail probability $\Pr[Q > b]$ in the corresponding infinite capacity model then provides a good approximation for the loss ratio [26]. Therefore, we calculate the tail probabilities and we achieve this by applying Darboux's theorem on $Q(z)$. To this end, we transform formula (6.25) for $\mathbf{Q}(z)$ into an expression for $Q(z)$ that is convenient to locate the dominant singularities. First, we rewrite (6.25) as follows:

$$\mathbf{Q}(z)\mathbf{r}_i(z) = \frac{g_i(z)}{[1-\lambda_i(z)][z^c - T_c(\lambda_i(z))]} \ ,$$

with $g_i(z)$ the right-hand-side of (6.25). As a next step, we sum both sides of this equation over i from 1 to N. On account of the distributive property of matrices, $\mathbf{R}(z)\mathbf{1} = \mathbf{1}$ and $\mathbf{Q}(z)\mathbf{1} = Q(z)$, we find

$$Q(z) = \sum_{i=1}^{N} \frac{g_i(z)}{[1-\lambda_i(z)][z^c - T_c(\lambda_i(z))]} \ .$$

In the sequel, we seek for the dominant singularities of $Q(z)$. First, recall that $|\lambda_1(z)| > |\lambda_i(z)|$ for all $i = 2, \ldots, N$ and for all $z \in]1, \Re_{\mathbf{A}}[$ ($\lambda_1(z)$ is the PF eigenvalue). Second, note that when $\tilde{z} \in]1, \Re_{\mathbf{A}}[$ is a zero of $[1-\lambda_j(z)]$, $g_j(z)$ also vanishes at $z = \tilde{z}$. Next, using similar arguments as in [1], where a continuous Markovian arrival process is considered, one can prove that $\lambda_1(z)$ is a strictly increasing and convex function for $z \in]1, \Re_{\mathbf{A}}[$. Hence, $z^c - T_c(\lambda_1(z))$ will have a unique solution in this region if $\lim_{z \uparrow \Re_{\mathbf{A}}} T_c(\lambda_1(z))/z^c > 1$, a requirement that we assume to be satisfied from now on. As a result, the dominant singularity of $Q(z)$, $\hat{z}$, is the zero from $z^c - T_c(\lambda_1(z))$ in $]1, \Re_{\mathbf{A}}[$. Taking these findings into account, we find that $Q(z)$ is in the neigbourhood of $\hat{z}$ proportional to

$$Q(z) \sim \frac{g_1(z)}{[1-\lambda_1(z)][z^c - T_c(\lambda_1(z))]} \ .$$

We thus find, by application of formula (1.4) of Darboux's theorem, the following approximation for $\Pr[Q > b]$:

$$\Pr[Q > b] \approx \frac{\hat{z}^{-(b+1)}}{1-\hat{z}} \frac{g_1(\hat{z})}{[1-\lambda_1(\hat{z})][c\hat{z}^{c-1} - T_c'(\lambda_1(\hat{z}))\lambda_1'(\hat{z})]} \ . \tag{6.31}$$

In practice, buffer dimensioning is an important assignment. For instance, one has to dimension the buffer so that the loss ratio is smaller than 10^{-6}. We can translate this problem to our setting: determine b such that $\Pr[Q > b] \leq 10^{-6}$. Taking the Neperian logarithm of this equation and on account of (6.31), we obtain:

$$b \geq \frac{6 \ln 10 + \ln K}{\ln \hat{z}} - 1 \ ,$$

with

$$K = \frac{1}{1-\hat{z}} \frac{g_1(\hat{z})}{[1-\lambda_1(\hat{z})][c\hat{z}^{c-1} - T_c'(\lambda_1(\hat{z}))\lambda_1'(\hat{z})]} \ .$$

Hence, the smallest buffer capacity $b \in \mathbb{N}$ that ensures a loss ratio not larger than 10^{-6} is equal to

$$b = \left\lceil \frac{6 \ln 10 + \ln K}{\ln \hat{z}} \right\rceil - 1 \ .$$

6.6 Numerical examples

In this section, we evaluate the influence of combining correlation in the arrival process and batch service on the behaviour of the system. To this end, we consider a numerical example whereby the number of background states N equals 2, and we assume that $p_{11} = p_{22}$. In view of the above assumptions, we define the coefficient of correlation γ between the states of two consecutive slots as

$$\gamma \triangleq \lim_{k \to \infty} \frac{\mathrm{E}[\tau_k \tau_{k+1}] - \mathrm{E}[\tau_k]\,\mathrm{E}[\tau_{k+1}]}{(\mathrm{Var}[\tau_k]\,\mathrm{Var}[\tau_{k+1}])^{1/2}} = 2p_{11} - 1 \ .$$

We also assume that no customers arrive when the background state equals 1 and that the number of arrivals in the other case is geometrically distributed, i.e. $A_1(z) = 1$ and $A_2(z) = 1/(1 + 2\lambda - 2\lambda z)$. We further consider a server of capacity 10 ($c = 10$). The service times are geometrically distributed with the mean length being dependent on the number of customers in the served batch. More specifically, the average time to serve a batch of n customers is equal to $8+0.2n$. Finally, the probability β that the server initiates a service even when less than l customers are present equals 0.05.

In Fig. 6.1, the mean system content $\mathrm{E}[U]$ is depicted versus the load ρ for several values of the correlation coefficient γ. It is assumed that the service threshold l equals 5. Fig. 6.1 demonstrates that positive correlation ($\gamma > 0$) leads to a significant larger $\mathrm{E}[U]$ as compared to the independent case ($\gamma = 0$). Hence, disregarding positive correlation can lead to a severe underrating of the mean system content. Fig. 6.1 also exhibits that ignoring negative correlation leads to some overestimation of $\mathrm{E}[U]$. We further perceive that these observations manifest themselves more as ρ increases. These conclusions are similar to those in multiserver systems with correlated arrivals (see e.g. [25], [60]).

Fig. 6.2 shows the tail probabilities $\Pr[Q > b]$ versus b in the case that the load ρ equals 0.6 and the service threshold l being 5. We perceive that positive correlation leads to much larger probabilities whereas negative correlation causes some smaller probabilities.

When we take a look at the buffer capacity required to ensure that the loss ratio is smaller than 10^{-6} (Fig. 6.3), we come to similar conclusions. Hence, we can state that correlation potentially has a huge impact on the buffer content.

Next, we investigate the server efficiency. Therefore, the filling degree - defined as $\mathrm{E}\left[\tilde{S}\right]/c$, the mean number of customers in a served batch divided by the server capacity - and the probability that the server processes a batch during a random slot are depicted versus the load in Fig. 6.4. We observe that positive correlation leads to a larger filling degree and a smaller serving probability, whereas the opposite holds (in a lesser degree) for negative correlation. Hence, positive correlation leads to a more efficient usage of the server. Indeed, in case of positive correlation, long periods exist during which the server is idle because no customers arrive. On the other hand, when customers arrive, this is likely to happen during many contiguous slots, so that the server then serves more customers.

As determining the optimal service threshold (we define it as the one that minimizes $\mathrm{E}[U]$) is of the utmost importance in batch-service systems, we study whether correlation affects this optimum. For this purpose, the optimal threshold is shown versus ρ in Fig. 6.5, for several values of γ. We perceive that the larger the correlation coefficient, the faster the optimum of l increases. Indeed, when, in the independent case, it becomes advantageous to postpone service until more customers have arrived, it can be beneficial in the positive correlated case to wait until even more customers have arrived, because when customers arrive it is very likely that other customers arrive in the subsequent slots.

We now investigate the impact of adopting the optimal threshold of the independent case in the correlated system. Therefore, we define the relative error as

$$\frac{\mathrm{E}\left[U\right]_{\tilde{l}_{\mathrm{opt}}} - \mathrm{E}\left[U\right]_{l_{\mathrm{opt}}}}{\left(\mathrm{E}\left[U\right]_{l_{\mathrm{opt}}} + \mathrm{E}\left[U\right]_{\tilde{l}_{\mathrm{opt}}}\right)/2} ,$$

with $\mathrm{E}\left[U\right]_{l_{\mathrm{opt}}}$ the mean system content in the correlated case when the optimal service threshold is adopted and $\mathrm{E}\left[U\right]_{\tilde{l}_{\mathrm{opt}}}$ the mean system content in the correlated system when the optimal threshold of the corresponding independent system is adopted. In Fig. 6.6, the relative errors are depicted for the example from Fig. 6.5. We observe that even when the optimal service threshold is different, the relative error is rather small. In view of this, the existing results of the corresponding independent system can be relied upon to determine a near-optimal service threshold. Adopting this near-optimal threshold has only a marginal impact on the mean system content.

Remark 31. *We have considered just one set of examples in this section. We have also examined additional examples (which we do not add to this dissertation), and the same conclusions could be drawn.*

Remark 32. *In the example, we have noticed that positive correlation has a larger effect on the behaviour of the system than negative correlation. We can explain this intuitively. Therefore, let us call state 1 the "inactive" (no arrivals) and state 2 the "active" state. When being in the inactive state, the temporary load (the load during several consecutive slots of the same state) becomes very small. This effect causes a temporarily smaller system content. When being in the active state, the temporary load becomes very large, which causes a temporarily larger system content. Due to the queueing effect (i.e. the system content increases exponentially for large load), the effect of being in active state is larger than the effect of being in inactive state. Of course, the longer active periods last (i.e. the larger the correlation), the more this effect plays a role.*

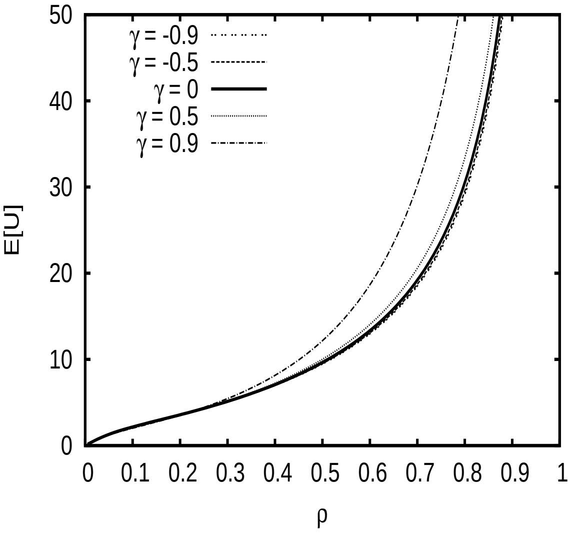


Figure 6.1: Mean system content $\mathrm{E}\left[U\right]$ versus the load ρ for several values of the correlation coefficient γ; $c = 10$, $l = 5$, $\beta = 0.05$, T_n geometrically distributed, $\mathrm{E}\left[T_n\right] = 8 + 0.2n$

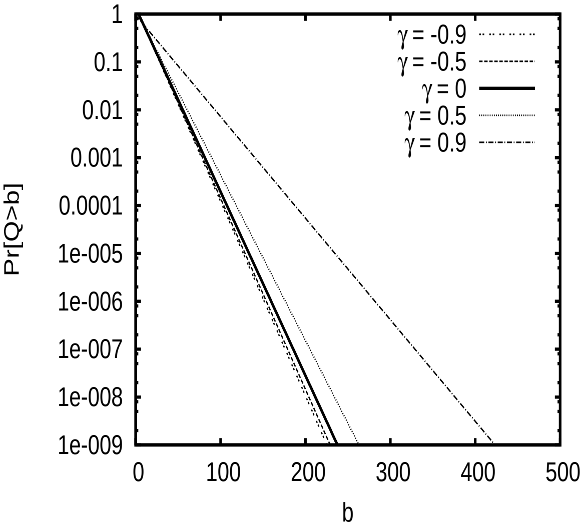


Figure 6.2: $\Pr\left[Q > b\right]$ versus b for several values of the correlation coefficient γ; $\rho = 0.6$, $c = 10$, $l = 5$, $\beta = 0.05$, T_n geometrically distributed, $\mathrm{E}\left[T_n\right] = 8 + 0.2n$

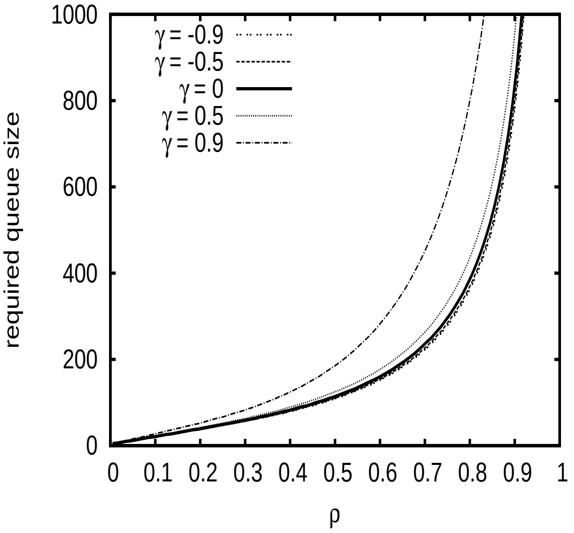
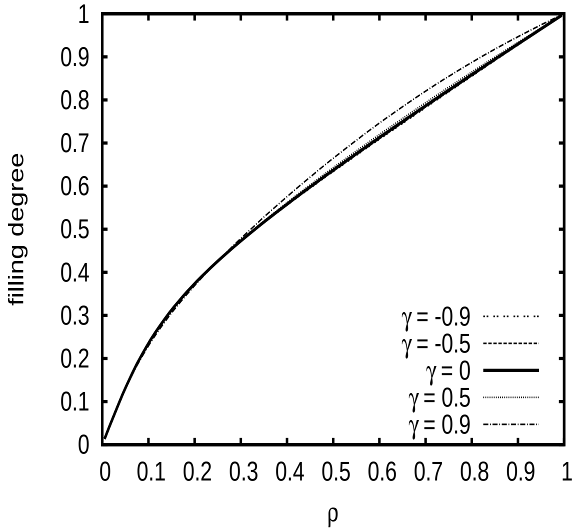
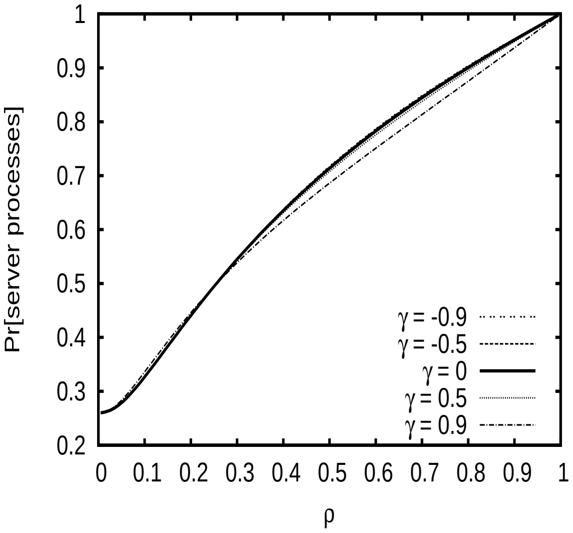


Figure 6.3: Required buffer capacity to ensure that the loss ratio is smaller than 10^{-6} versus the load ρ for several values of the correlation coefficient γ; $c = 10$, $l = 5$, $\beta = 0.05$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$



(a) filling degree



(b) probability that the server processes a batch during a random slot

Figure 6.4: Server efficiency versus the load ρ for several values of the correlation coefficient γ; $c = 10$, $l = 5$, $\beta = 0.05$, T_n geometrically distributed, $\mathrm{E}\left[T_n\right] = 8 + 0.2n$

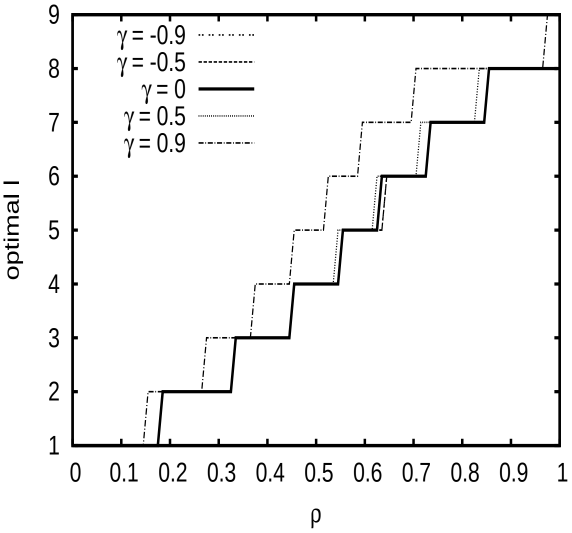


Figure 6.5: Optimal service threshold l versus the load ρ for several values of the correlation coefficient γ; $c = 10$, $\beta = 0.05$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$

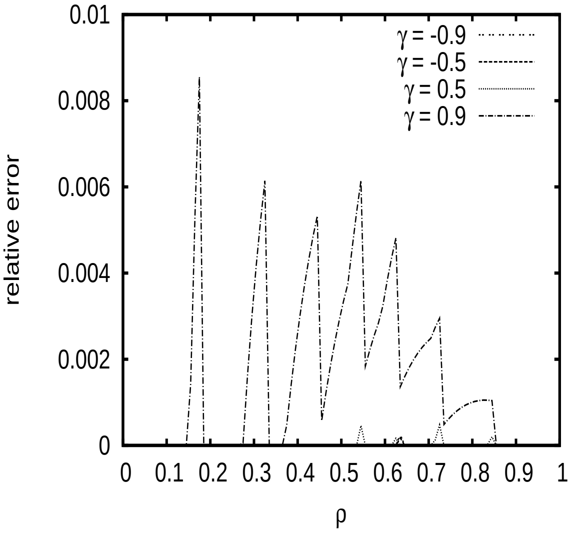


Figure 6.6: Relative error versus the load ρ for several values of the correlation coefficient γ; $c = 10$, $\beta = 0.05$, T_n geometrically distributed, $\mathrm{E}[T_n] = 8 + 0.2n$

Chapter 7

Conclusions

Throughout this dissertation, we have investigated various aspects of a versatile discrete-time queueing model with batch service. The model includes batch arrivals, meaning that several customers can arrive during a time slot, and a service threshold for the minimum number of customers that have to be present before a new service can be initiated in combination with a timer mechanism to avoid that this threshold strategy leads to excessive delays. In addition, the service time of a batch has a general distribution and is dependent on the number of customers within it. We now summarize the main contributions for every chapter separately.

Chapter 2

In chapter 2, we have deduced a fundamental formula from which a spectrum of known as well as novel quantities related to the buffer content have been extracted. We have thereby adopted an approach based on probability generating functions (PGFs) and supplementary variables. We have also demonstrated that our results can be applied to select an efficient service threshold and timer in real-life batch-service systems. We have also demonstrated this in detail for the group-screening application.

Chapter 3

Obtaining performance measures for batch-service queueing systems often requires additional numerical work, namely the calculation of c zeroes (c is the server capacity) and the solution of a set of c equations. These numerical calculations can become unfeasible when the server capacity c becomes large. For this reason and also for interpretational motives, we have developed, in chapter 3, light- and heavy-traffic approximations for the quantities from chapter 2. In order to deduce light-traffic approximations, we have expanded all quantities

as a Taylor series about $\lambda = 0$ (λ is the mean arrival rate), whereas we have taken the limit $\lambda \to c/\mathrm{E}[T_c]$ (thus letting the load going to one) for the heavy-traffic approximations, where $\mathrm{E}[T_c]$ represents the mean service time of a batch containing c customers. The resulting light-traffic approximations reduce the amount of numerical work considerably, in the sense that no zeroes have to be calculated anymore, and the heavy-traffic approximations require no numerical work at all.
To the best of our knowledge, light- and heavy-traffic approximations have not been studied before for batch-service queueing models.

Chapter 4

Whereas the buffer content has been covered in chapters 2 and 3, we have turned our focus in chapters 4 and 5 to the customer delay. One of the main contributions of this dissertation is that we combine the study of the customer delay in a batch-service queueing model with batch arrivals, while in existing literature the customer delay is only investigated in the case where customers arrive individually. In chapter 4, we have deduced moments, and in chapter 5, we have developed tail probabilities of the customer delay. In order to calculate moments, we have adopted an approach based on PGFs and supplementary variables, and we have subdivided the delay of a random customer as the sum of two components: the time to serve previously arrived batches (the queueing delay) and the time, starting at the end of the queueing delay, required to fill the batch with enough customers or until the service timer expires (postponing delay).

Chapter 5

The approach from chapter 4 fits well for obtaining moments, but not for calculating tail probabilities of the customer delay. We have therefore adopted a different method in chapter 5. We have redefined the postponing delay, so that the customer delay becomes the maximum of both components instead of the sum. This approach has eventually led us to approximations for the tail probabilities. These formulas can for instance be applied to evaluate the performance of real-time applications, where the quality of service (QoS) is typically expressed in terms of the order of magnitude of the probability that a real-time packet experiences an excessive delay.

Chapter 6

In nearly all batch-service queueing systems covered in literature, customer arrivals occur independently. In many real-life circumstances however, this is unrealistic. For instance, in modern telecommunication systems, a traffic source which is inactive in a given time slot is very likely to remain inactive for a long time. We have therefore studied in chapter 6 the buffer content in a batch-service model that includes dependent arrivals. We have modeled the customer arrivals by a discrete-batch Markovian arrival process (D-BMAP). Our work differs in terms of the model under investigation as well as the calculated quantities from the few papers that also consider this topic. In addition, we have evaluated more thoroughly the influence of correlation on the behaviour of batch-service queueing systems and more specifically, we have investigated the influence on the optimal service threshold. We have shown that correlation merely has a small impact on the service threshold that minimizes the buffer content, and consequently, that the existing results of the corresponding independent system can be applied to determine a near-optimal service threshold, which is an important finding for practitioners. On the other hand, we have demonstrated that for other purposes, such as performance evaluation and buffer management, correlation in the arrival process cannot be ignored, a conclusion that runs along the same lines as in queueing models without batch service.

Summarized, we believe that we have deduced a large spectrum of performance measures for a rather versatile discrete-time queueing model with batch service. These performance measures are useful tools to evaluate real-life batch-service queueing systems, such as group-screening facilities, various production and transport processes, telecommunication systems where packets are transmitted in bursts, et cetera.

Finally, we would like to close by mentioning some interesting challenges, which naturally ensue from our dissertation:

- In chapter 6, we have included a correlated arrival process in our model and we have studied the buffer content. The customer delay in this case has not been covered up to now and deserves further investigation.

- Analogously, in chapter 3, light-traffic approximations have been established for the buffer content. Obtaining such approximations for the customer delay would be an interesting research topic.

- Also, deducing approximations in case of intermediate values of the load is a tempting direction for future research. A possible approach would be to interpolate based on the obtained light- and heavy-traffic formulas.

- The light-traffic formulas from chapter 3 eliminate the calculation of zeroes of $z^c - T_c(A(z))$, but it is still required to solve sets of equations. In the future, we intend to elaborate on the obtained formulas in order to obtain closed-form formulas. First research in that direction suggests that the set of equations for $d_1(n), 1 \leq n \leq c-1$ in (3.24) can, with some manipulations, be transformed into a set of equations whose coefficient matrix is a Vandermonde matrix. As a result, these equations can then be solved explicitly. Of course, this is only the case for $\beta \neq 0$. For $\beta = 0$, we are investigating whether a similar approach would work.

- In chapters 4 and 5, we have studied the time that a customer spends in the queue. In a classical queueing system, the time a customer spends in the queue is independent of its service time, so that the PGF of the time the customer spends in the entire system (queue and server) is the product of both individual PGFs. As in our model the service time of a tagged customer depends on the number of customers served in the batch containing the tagged customer, the service time is dependent on the time the tagged customer spends in the queue. This issue leads to considerable complications. For instance, we do not only have to calculate the time until the server is allowed to initiate a new service, but we also have to compute the number of customers at that moment. As this asks for a distinct approach, we leave this issue for future research.

- Although, in theory, the number of background states of the underlying Markov chain of the D-BMAP considered in chapter 6, can be large, spectral decomposition becomes unfeasible in that case. Therefore, a refinement of the results that have been reported would be welcomed.

- The probability β to start a new service when less customers than the service threshold are present is independent of the number of available customers. Inclusion of a dependency could be interesting from practical point of view and would especially be challenging when studying the customer delay. The reason is that relation (4.2) does not hold anymore, exactly due to the dependency of the number of present customers on the timer parameter.

- Throughout this dissertation, we have considered various numerical examples, wherein we have studied the influence of the system parameters on the overall performance. We have always depicted various performance measures as a function of one parameter, while keeping the other parameters constant. However, when an optimal service strategy has to be determined in practice, one has to select the best “combination” of decision parameters l (service threshold) and β (timer parameter). In order to find the optimal combination, we can combine techniques from optimization theory with the performance measures deduced throughout this dissertation. This topic might be a direction for future research.

- We intend to investigate queueing systems with multiple batch servers. As multi-service queueing models with generally distributed service times are extremely hard to analyze, we will consider simple service-time distributions (deterministic, geometric) as the combination with batch service is even more difficult.

Appendix A

Analyticity at $\lambda = 0$

In this appendix, we show that, if $A(\lambda, z)$ is analytic in

$$\mathcal{D} = \{(\lambda, z) : |\lambda| < \delta, |z| < 1 + \gamma\} \ , \qquad \delta > 0, \gamma > 0 \ ,$$

then (i) the zeroes $z_i(\lambda)$, $i = 0, \ldots, c-1$ of $z^c - T_c(A(\lambda, z))$ are analytic at $\lambda = 0$, (ii) the unknown probabilities $d(\lambda, n)$ are analytic at $\lambda = 0$ and (iii) all quantities related to the buffer content (for instance $U(\lambda, z)$) are analytic at $\lambda = 0$ for $|z| \leq 1$.

From $A(\lambda, z)$ being analytic in $\mathcal{D}$, it follows that $f(\lambda, z) \triangleq z^c - T_c(A(\lambda, z))$ is analytic in $\mathcal{D}$ (mark that we have previously assumed that the radius of convergence of $T_c(z)$ is larger than 1). Hence, $f(\lambda, z)$ is analytic in a neighbourhood of the points $(0, \varepsilon_i)$, $i = 0, \ldots, c-1$ (a). Further, $f(0, \varepsilon_i) = 0$ and

$$\frac{\partial}{\partial z} f(\lambda, z)\bigg|_{\lambda=0, z=\varepsilon_i} = \frac{c}{\varepsilon_i} \neq 0 \qquad \text{(b)} \ .$$

From (a) and (b) and the implicit function theorem, it follows that there exists a unique function $z_i(\lambda)$, that satisfies

$$f(\lambda, z_i(\lambda)) = 0 \ ,$$

$$z_i(0) = \varepsilon_i \ ,$$

and that is analytic at $\lambda = 0$ for $i = 0, \ldots, c-1$. Next, it is possible to prove that (i) implies (ii) by virtue of the implicit function theorem (see e.g. [69]). Finally, from the calculus of analytic functions, it follows that (iii) also holds.

Bibliography

[1] Abate J., Choudhry G.L., Whitt W. (1994), Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models*, 10(1), 99–143.

[2] Abolnikov L., Dukhovny A. (2003), Optimization in HIV screening problems. *Journal of Applied Mathematics and Stochastic Analysis*, 16(4), 361–374.

[3] Adan I.J.B.F., Resing J.A.C. (2000), Multi-server batch-service systems. *Statistica Neerlandica*, 54(2), 202–220.

[4] Adan I.J.B.F., Kulkarni V.G. (2003), Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Systems*, 45, 113–134.

[5] Adan I.J.B.F., van Leeuwaarden J.S.H., Winands E.M.M. (2006), On the application of Rouché's theorem in queueing theory, *Operations Research Letters*, 34, 355–360.

[6] Alfa A.S. (1995), A discrete MAP/PH/1 queue with vacations and exhaustive time-limited service. *Operations Research Letters*, 18, 31-40.

[7] Arumuganathan R., Jeyakumar S. (2005), Steady state analysis of a bulk queue with multiple vacations, setup times with N-policy and closedown times. *Applied Mathematical Modelling*, 29(10), 972–986.

[8] Asmussen S. (1992), Light traffic equivalence in single server queues. *Annals of Applied Probability*, 2(3), 555–574.

[9] Bailey N.T.J. (1954), On queueing processes with bulk service. *Journal of the Royal Statistical Society*, 16(1), 80–87.

[10] Baiocchi A. (1994), Analysis of the loss probability in MAP/G/1/K-queues Part I: Asymptotic theory. *Stochastic Models*, 10, 867-893.

[11] Banik A.D., Gupta U.C., Pathak S.S. (2006). BMAP/G/1/N queue with vacations and limited service discipline. *Applied Mathematics and Computations*, 180, 707–721.

[12] Banik A.D. (2009), Queueing analysis and optimal control of BMAP/$G^{(a,b)}$/1/N and BMAP/$MSP^{(a,b)}$/1/N systems. *Computers and Industrial Engineering*, 57, 748–761.

[13] Bar-Lev, S.K., Stadje, W., Van der Duyn Schouten, F.A. (2004), Optimal Group Testing with Processing Times and Incomplete Identification. *Methodology and Computing in Applied Probability*, 6, 55–72.

[14] Bar-Lev S.K., Stadje W., Van der Duyn Schouten F.A. (2005), Multinomial group testing models with incomplete identification. *Journal of Statistical Planning and Inference*, 135, 384–401.

[15] Bar-Lev S.K., Stadje W., Van der Duyn Schouten F.A. (2006), Group testing procedures with incomplete identification and unreliable testing results. *Applied Stochastic Models in Business and Industry*, 22, 281–296.

[16] Bar-Lev S.K., Parlar M., Perry D., Stadje W., Van der Duyn Schouten F.A. (2007), Applications of bulk service queues to group testing models with incomplete identification. *European Journal of Operational Research*, 183, 226–2374.

[17] Behets F., Bertozzi S., Kasali M., Kashamuka M., Atikala L., Brown C., Ryder R.W., Quinn T.C. (1990), Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost-efficiency models. *AIDS*, 4(8), 737–741.

[18] Bellalta, B. (2009), A queueing model for the non-continuous frame assembly scheme in finite buffers. *Proceedings of the 16th international conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2009)*, Madrid, June 9-12, 219–233.

[19] Benes V. (1965), *Mathematical theory of connecting networks and telephone traffic*. Academic press, New York.

[20] Blaszczyszyn B., Rolski T. (1993), Queues in series in light traffic. *Annals of Applied Probability*, 3(3), 881–896.

[21] Blondia C. (1991), Finite capacity vacation models with non-renewal input. *Journal of Applied Probability*, 28, 174-197.

[22] Blondia C. (1993), A discrete-time batch Markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science*, 32, 3–23.

[23] Bruin A., Rossum A., Visser M., Koole G. (2007), Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2), 125–137.

[24] Bruneel H. (1983), Buffers with stochastic output interruptions. *Electronics Letters*, 19, 461–463.

[25] Bruneel H. (1988), Queueing behavior of statistical multiplexers with correlated inputs. *IEEE Transactions on Communications*, 36(12), 1339–1341.

[26] Bruneel H., Steyaert B., Desmet E., Petit G.H. (1992), Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. *European Journal of Operational Research*, 76, 563–572.

[27] Buzacott J.A., Yao D.D. (1986), On queueing network models of flexible manufacturing systems. *Queueing Systems: Theory and Applications*, 1(1), 5–27.

[28] Chakravarthy S. (1992), A finite-capacity GI/PH/1 queue with group services. *Naval Research Logistics*, 39(3), 345–357.

[29] Chang S.H., Choi D.W. (2005), Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations. *Computers and Operations Research*, 32, 2213–2234.

[30] Chang S.H., Takine T. (2005), Factorization and stochastic decomposition properties in bulk queues with generalized vacations. *Queueing Systems*, 50, 165–183.

[31] Chaudhry M.L. (1965), Correlated queueing. *CORS Journal*, 3, 142-151.

[32] Chaudhry M.L., Templeton J.G.C. (1983), *A first course in bulk queues.* John Wiley & Sons.

[33] Chaudhry M.L., Gupta U.C. (1999), Modelling and analysis of M/$G^{a,b}$/1/N queue A simple alternative approach. *Queueing Systems*, 31, 95-100.

[34] Chaudhry M.L., Gupta U.C. (2003), Analysis of a finite-buffer bulk-service queue with discrete-Markovian arrival process: D-MAP/$G^{a,b}$/1/N. *Naval Research Logistics*, 50(4), 345–363.

[35] Chaudhry M.L., Gupta U.C. (2003), Queue length distributions at various epochs in discrete-time D-MAP/G/1/N queue and their numerical evaluation. *International Journal of Information and Management Sciences*, 14(3), 67–83.

[36] Chen Y., Qiao C., Yu X. (2004), Optical burst switching (OBS): a new area in optical networking research. *IEEE Network*, 18(3), 16–23.

[37] Claeys D., Walraevens J., Laevens K., Bruneel H. (2007), A discrete-time queueing model with a batch server operating under the minimum batch size rule. *Lecture Notes in Computer Science*, 4712, 248–259.

[38] Claeys D., Walraevens J., Laevens K., Bruneel H. (2007), A batch server with service times dependent on the number of served customers. *Proceedings of the European Simulation and Modelling Conference (ESM 2007), St. Julians, October 22-24*, 214–219.

[39] Claeys D., Laevens K., Walraevens J., Bruneel H. (2008), Delay in a discrete-time queueing model with batch arrivals and batch services. *Proceedings of the Information Technology: New Generations Conference (ITNG 2008), Las Vegas, Nevada, April 7-9*, 1040–1045.

[40] Claeys D., Laevens K., Walraevens J., Bruneel H. (2010), Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service. *Mathematical Methods of Operations Research*, 72(1), 1–23.

[41] Claeys D., Walraevens J., Laevens K., Bruneel H. (2010), Delay analysis of two batch-service queueing models with batch arrivals: $Geo^X/Geo^c/1$. *4OR*, 8(3), 255–269.

[42] Claeys D., Walraevens J., Laevens K., Bruneel H. (2010), A queueing model for general group screening policies and dynamic item arrivals. *European Journal of Operational Research*, 207(2), 827–835.

[43] Claeys D., Walraevens J., Laevens K., Bruneel H. (2011), Analysis of threshold-based batch-service queueing systems with batch arrivals and general service times. *Performance Evaluation*, 68(6), 528–549.

[44] Claeys D., Steyaert B., Walraevens J., Laevens K., Bruneel H., Tail distribution of the delay in a general batch-service queueing model. Submitted to *Computers and Operations Research.*

[45] Claeys D., Walraevens J., Laevens K., Steyaert B., Bruneel H. (2010), A batch-service queueing model with a discrete batch Markovian arrival process. *Proceedings of the 17th international conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2010)*, Cardiff, June 14-16, 1–13.

[46] Claeys D., Steyaert B., Walraevens J., Laevens K., Bruneel H., Analysis of a versatile batch-service queueing model with correlation in the arrival process. Submitted to *Performance Evaluation.*

[47] Daley D.J., Rolski T. (1991), Light traffic approximation in queues. *Mathematics of Operations Research*, 16, 57–71.

[48] De Turck K., De Vuyst S., Fiems D., Wittevrongel S. (2008), Performance analysis of the IEEE 802.16e sleep mode for correlated downlink traffic. *Telecommunication Systems*, 39, 145–156.

[49] Dorfman R. (1943), The detection of defective members of large populations. *Annals of Mathematics and Statistics*, 14, 436–440.

[50] Doshi B.T. (1986), Queueing systems with vacations - a survey. *Queueing Systems*, 1, 29–66.

[51] Downton F. (1955), Waiting time in bulk service queues. *Journal of the Royal Statistical Society, Series B (Methodological)*, 17(2), 256–261.

[52] Du D-Z., Hwang F.K. (2000), *Combinatorial group testing and its applications.* World Scientific.

[53] Emmanuel J.C., Bassett M.T., Smith H.J., Jacobs J.A. (1988), Pooling of sera for human immunodeficiency virus (HIV) testing: an economical method for use in developing countries. *Journal of Clinical Pathology*, 41, 582–585.

[54] Erlang A.K. (1925), Calcul des probabilités et conversations téléphoniques. *Revue Générale de L'électricité*, 18(8), 305–309.

[55] Ferdinand A.E. (1971), An analysis of the machine interference model. *IBM Systems Journal*, 2, 129–142.

[56] Ferrandiz J.M. (1993), The BMAP/GI/1 queue with server set-up times and server vacations. *Advances in Applied Probability*, 25(1), 235-254.

[57] Fiems D., Bruneel H. (2002), A note on the discretization of Little's result. *Operations Research Letters*, 30, 17–18.

[58] Fiems D., Bruneel H. (2002), Analysis of a discrete-time queueing system with timed vacation. *Queueing Systems*, 42(3), 243–254.

[59] Frigui I., Alfa A.S., Xu X. (1997), Algorithms for computing waiting time distributions under different queue disciplines for the D-BMAP/PH/1. *Naval Research Logistics*, 44, 559-576.

[60] Gao P., Wittevrongel S., Bruneel H. (2004), On the behavior of multiserver buffers with geometric service times and bursty input traffic. *IEICE Transactions on Commununications*, 12, 3576–3583.

[61] Gazis D.C. (2002), The origins of traffic theory. *Operations Research*, 50(1), 69–77.

[62] Gong W.B., Hu J.Q. (1992), The MacLaurin series for the GI/G/1 queue. *Journal of Applied Probability*, 29, 176–184.

[63] Goswami V., Mohanty J.R., Samanta S.K. (2006), Discrete-time bulk-service queues with accessible and non-accessible batches. *Applied Mathematics and Computation*, 182(1), 898–906.

[64] Gupta U.C., Laxmu P.V. (2001), Analysis of the MAP/$G^{a,b}$/1/N queue. *Queueing Systems*, 38, 109–124.

[65] Gupta U.C., Goswami V. (2002), Performance analysis of finite buffer discrete-time queue with bulk service. *Computers and Operations Research*, 29, 1331–1341.

[66] Gupta U.C., Sikdar K. (2004), A finite capacity bulk service queue with single vacation and Markovian arrival process. *Journal of Applied Mathematics and Stochastic Analysis*, 2004(4), 337–357.

[67] Gupta U.C., Samanta S.K., Sharma R.K., Chaudhry M.L. (2007), Discrete-time single-server finite buffer queues under discrete Markovian arrival process with vacations. *Performance Evaluation*, 64, 1–19.

[68] Herrmann C. (2001), The complete analysis of the discrete time finite DBMAP/G/1/N queue. *Performance Evaluation*, 43, 95-121.

[69] Hooghiemstra G., Keane M., Van De Ree S. (1988), Power series for stationary distributions of coupled processor models. *Siam Journal of Applied Mathematics*, 48(5), 1159–1166.

[70] Janssen A.J.E.M., van Leeuwaarden J.S.H. (2005), Analytic computation schemes for the discrete-time bulk service queue. *Queueing Systems*, 50, 141–163.

[71] Kasahara S., Takine T., Takahashi Y., Hasegawa T. (1996), MAP/G/1 queues under N-policy with and without vacations. *Journal of Operational Research Society Japan*, 39(2), 188-212.

[72] Kendall D.G. (1953), Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *Annals of Mathematical Statistics*, 24(3), 338–354.

[73] Kienzle M.G., Sevcik K.C. (1979), Survey of analytic queueing network models of computer systems. *ACM Sigmetrics Performance Evaluation Review*, 8(3), 113–129.

[74] Kim N.K., Chae K.C., Chaudhry M.L. (2004), An invariance relation and a unified method to derive stationary queue lengths. *Operations Research*, 52(5), 756–764.

[75] Kim N.K., Chaudhry M.L. (2006), Equivalences of batch-service queues and multi-server queues and their complete simple solutions in terms of roots. *Stochastic Analysis and Applications*, 24(4), 753–766.

[76] Kim B., Kim J. (2010), Queue size distribution in a discrete-time D-BMAP/G/1 retrial queue. *Computers and Operations Research*, 37(7), 1220–1227.

[77] Lee H.W., Moon J.M., Kim B.K., Park J.G., Lee S.W. (2005), A simple eigenvalue method for low-order D-BMAP/G/1 queues. *Mathematical Modelling*, 29, 277-288.

[78] Lee H.W., Ahn B.Y., Park N.I. (2001), Decompositions of the queue length distributions in the MAP/G/1 queue under multiple and single vacations with N-policy. *Stochastic Models*, 17(2), 157-190.

[79] Little J.D.C. (1961), A proof of the queueing formula $L = \lambda W$. *Operations Research*, 9, 383–387.

[80] Lu K., Wu D., Fang Y., Qiu R.C. (2005), Performance analysis of a burst-frame-based MAC Protocol for ultra-wideband ad hoc networks. *Proceedings of the IEEE International Conference on Communications (ICC 2005)*, Seoul, May 16-20, Vol.5, 2937–2941.

[81] Lucantoni D.M., Meier-Hellstern K.S., Neuts M.F. (1990), A single-server queue with server vacations and a class of non-renewal process. *Advances in Applied Probability*, 22, 676-705.

[82] Macula A.J. (1999), Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening. *Journal of Combinatorial Optimization*, 2, 385–397.

[83] Macula, A.J. (1999), Probabilistic nonadaptive group testing in the presence of errors and DNA library screening. *Annals of Combinatorics*, 3, 61–69.

[84] Mahnke R., Kaupuzs J. (2001), Probabilistic description of traffic flow. *Networks and Spatial Economics*, 1(1-2), 103–136.

[85] Matendo S.K. (1993), A single-server queue with server vacations and a batch Markovian arrival process. *Cahiers Centre Etudes Rech. Oper.*, 35(1-2), 87-114.

[86] Matendo S.K. (1994), Some performance measures for vacation models with a batch Markovian arrival process. *Journal of Applied Mathematics and Stochastic Analysis*, 7(2), 111-124.

[87] Masuyama H. (2003), *Studies on algorithmic analysis of queues with batch Markovian arrival streams.* Phd dissertation, Kyoto, Japan.

[88] Medhi J. (1975), Waiting time distributions in a Poisson queue with a general bulk service rule. *Management Science*, 21(2), 777–782.

[89] Meyer C. (2000), *Matrix analysis and applied linear algebra.* Society for Industrial and Applied Mathematics, Philadelphia.

[90] Miller R.G. (1959), A contribution to the theory of bulk queues. *Journal of the Royal Statistical Society Series B (Methodological)*, 21(2), 320–337.

[91] Neuts M.F. (1967), A general class of bulk queues with Poisson input. *Annals of Mathematical Statistics*, 38, 759–770.

[92] Niu Z., Takahashi Y. (1999), A finite-capacity queue with exhaustive vacation/close-down/setup times and Markovian arrival processes. *Queueing Systems Theory and Applications*, 31(1-2), 1-23.

[93] Niu Z., Shu T., Takahashi Y. (2003), A vacation queue with setup and close-down times and batch Markovian arrival processes. *Performance Evaluation*, 54, 225–248.

[94] Powell W.B., Humblet P. (1986), The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure. *Operations Research*, 34(2), 267–275.

[95] Preater J. (2001), A bibliography of queues in health and medicine. *Keele Mathematics Research Report.*

[96] Qiao C.M., Yoo M.S. (1999), Optical burst switching (OBS) - a new paradigm for an optical Internet. *Journal of High Speed Networks*, 8(1), 69–84.

[97] Reiman M., Simon B. (1989), Open queueing systems in light traffic. *Mathematics of Operations Research*, 14(1), 26–59.

[98] Samanta S.K., Chaudhry M.L., Gupta U.C. (2007), Discrete-time $Geo^X|G^{(a,b)}|1|N$ queues with single and multiple vacations. *Mathematical and Computer Modelling*, 45, 93–108.

[99] Samanta S.K., Gupta U.C., Sharma R.K. (2007), Analyzing discrete-time D-BMAP/G/1/N queue with single and multiple vacations. *European Journal of Operational Research*, 182(1), 321–339.

[100] Schellhaas H. (1994), Single server queues with a batch Markovian arrival process and server vacations. *OR Spektrum*, 15, 189-196.

[101] Sigman K. (1992), Light traffic for workload in queues. *Queueing Systems*, 11, 429–442.

[102] Shioyama T., Kise H. (1989), Optimization in production systems a survey on queuing approaches. *Journal of the Operations Research Society of Japan*, 32(1), 34–55.

[103] Sikdar K., Gupta U.C. (2005), Analytic and numerical aspects of batch service queues with single vacation. *Computers and Operations Research*, 32, 943–966.

[104] Sikdar K., Gupta U.C. (2005), The queue length distributions in the finite buffer bulk-service MAP/G/1 queue with multiple vacations. *Sociedad de Estadistica e Investigacidn Operativa Top*, 13(1), 75–103.

[105] Sobel M., Groll P.A. (1959), Group testing to eliminate efficiently all defectives in a binomial sample. *Bell Systems Technical Journal*, 28, 1179-1252.

[106] Steyaert B. (2008), Analysis of generic discrete-time buffer models with irregular packet arrival patterns. http://biblio.ugent.be/record/471285, Phd-thesis, Ghent University (promoter: Herwig Bruneel)

[107] Steyaert B., Walraevens J., Fiems D., De Vleeschauwer D., Bruneel, H. (2008), Heterogeneous sources model for DSL access multiplexers. *Electronics Letters*, 44(21), 1282–1283.

[108] Tian N., Zhang G.H. (2006), *Vacation queueing models.* International series in operations research & management science, 93.

[109] Tu X.M., Litvak E., Pagano M. (1995), On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika*, 82, 287–297.

[110] Van der Duyn Schouten F.A., Vanneste S.G. (1995). Maintenance optimization of a production system with buffer capacity. *European Journal of Operational Research*, 82, 323–338.

[111] Van Houdt B., Lenin R.B., Blondia C. (2003), Delay distribution of (im)patient customers in a discrete time D-MAP/PH/1 queue with age-dependent service times. *Queueing Systems*, 45, 59–73.

[112] Van Velthoven J., Van Houdt B., Blondia C. (2005), Response time distribution in a D-MAP/PH/1 queue with general customer impatience. *Stochastic Models*, 21, 745–765.

[113] Wein L.M., Zenios S.A. (1996), Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research*, 44, 543–569.

[114] Xie M., Tatshuoka K., Sacks J., Young S. (2001), Group testing with blockers and synergism. *Journal of the American Statistical Association*, 96, 92–102.

[115] Yi X.W., Kim N.K., Yoon B.K., Chae K.C. (2007), Analysis of the queue-length distribution for the discrete-time batch-service $Geo^X|G^{a,Y}|1|K$ queue. *European Journal of Operational Research*, 181, 787–792.

[116] Zhang Z. (1991), Analysis of a discrete-time queue with integrated bursty inputs in ATM networks. *International Journal on Digital and Analog Communication Systems*, 4, 191–203.

[117] Zhao Y.Q., Campbell L.L. (1996), Equilibrium probability calculations for a discrete-time bulk queue model. *Queueing Systems*, 22, 189–198.

[118] Zhu L., Hughes-Oliver J., Young S. (2001), Statistical decoding of potent pools based on chemical structure. *Biometrics*, 57, 922–930.