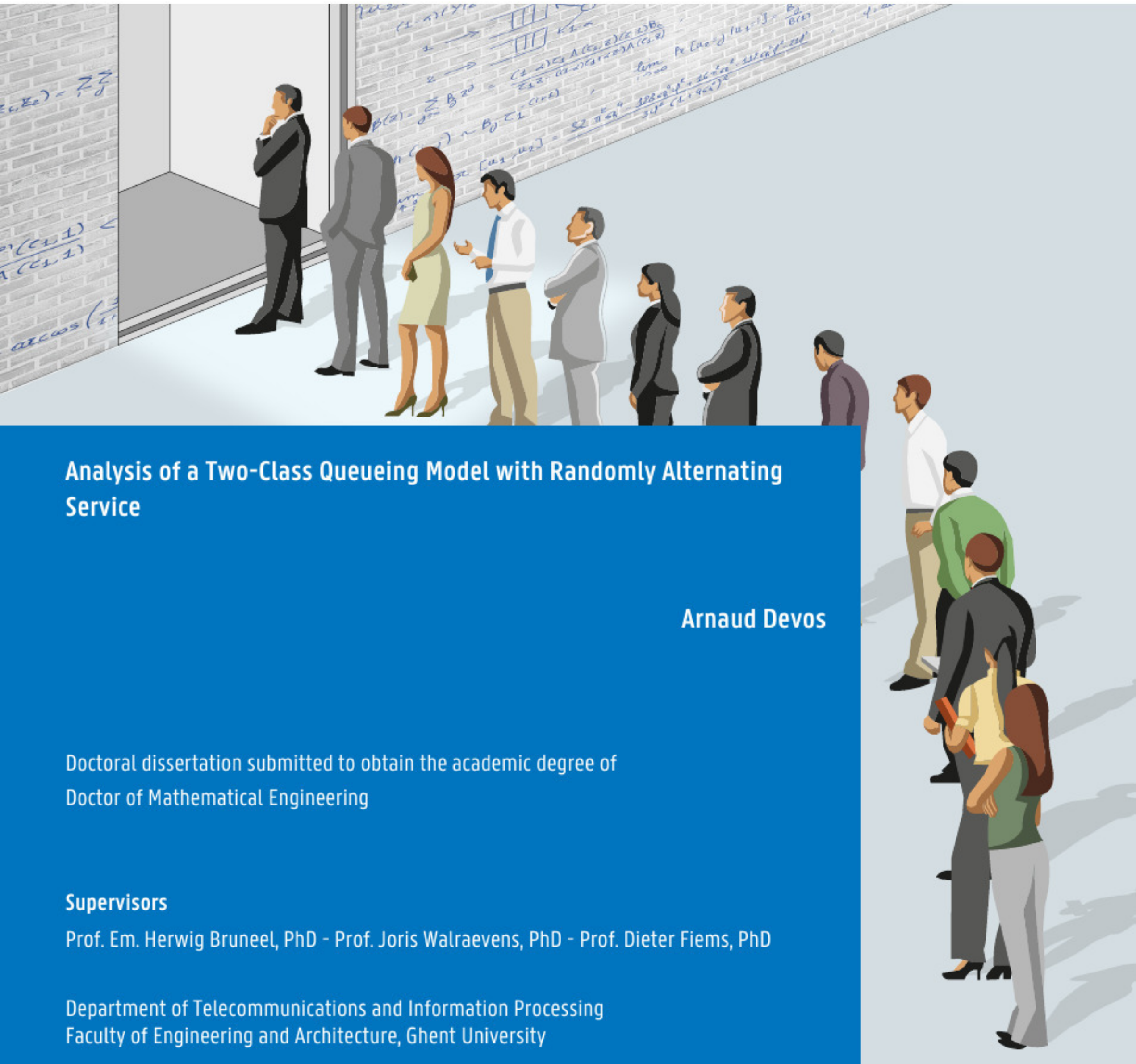
FACULTY OF ENGINEERING
AND ARCHITECTURE



Analysis of a Two-Class Queueing Model with Randomly Alternating Service

Arnaud Devos

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Mathematical Engineering

Supervisors

Prof. Em. Herwig Bruneel, PhD - Prof. Joris Walraevens, PhD - Prof. Dieter Fiems, PhD

Department of Telecommunications and Information Processing
Faculty of Engineering and Architecture, Ghent University

May 2022

GHENT
UNIVERSITY

FACULTY OF ENGINEERING
AND ARCHITECTURE

Analysis of a Two-Class Queueing Model with Randomly Alternating Service

Arnaud Devos

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Mathematical Engineering

Supervisors
Prof. Em. Herwig Bruneel, PhD - Prof. Joris Walraevens, PhD - Prof. Dieter Fiems, PhD

Department of Telecommunications and Information Processing
Faculty of Engineering and Architecture, Ghent University

May 2022

GHENT
UNIVERSITY

ISBN 978-94-6355-589-0
NUR 919, 992
Wettelijk depot: D/2022/10.500/30

Members of the Examination Board

Chair

Prof. Hennie De Schepper, PhD, Ghent University

Other members entitled to vote

Prof. Ioannis Dimitriou, PhD, University of Patras, Greece
Prof. Stella Kapodistria, PhD, Technische Universiteit Eindhoven, the Netherlands
Prof. Michèle Vanmaele, PhD, Ghent University
Prof. Sabine Wittevrongel, PhD, Ghent University

Supervisors

Prof. Em. Herwig Bruneel, PhD, Ghent University
Prof. Joris Walraevens, PhD, Ghent University
Prof. Dieter Fiems, PhD, Ghent University

Dankwoord

De voorbije jaren heb ik het geluk gehad om onderzoek te mogen verrichten in een boeiend domein als dat van de wachtlijntheorie. Met trots kan ik enkele van mijn resultaten tonen in dit proefschrift. Maar eerst wens ik graag mijn dankbaarheid voor enkele personen te vereeuwigen door middel van dit dankwoord. *Here we go.*

Vooreerst wens ik mijn drie promotoren te bedanken. Het was heel fijn -en belangrijk- voor mij dat jullie zo in mij geloofden. Ik kon altijd bij jullie terecht voor vragen, over wat dan ook.

Herwig, bedankt voor deze mooie kans die je mij gegeven hebt. Als ik het goed heb, ben ik de laatste doctoraatsstudent die gestart is onder jouw promotorschap. Ik vond het altijd zo boeiend om naar je uitleg te luisteren, gaande van wachtlijntheorie, de werking binnen de universiteit tot de geschiedenis van SMACS. Ik zal altijd met plezier terugdenken aan deze gesprekken. We zijn er dan wel niet in geslaagd om het boundary-value probleem volledig op te lossen zoals we voor ogen hadden, stiekem hoop ik dat je toch trots bent op wat we hebben bereikt.

Joris, volgens mij kan ik je niet genoeg bedanken. Je stond werkelijk altijd klaar om te luisteren naar mijn vragen, mijn werk te verbeteren (en dan nog wel in zeer korte tijd) of om mij gerust te stellen als ik dat nodig had. De momenten waarop we ons bogen over de laatste wiskundige details in een bewijs of berekening, heb ik altijd als één van de meest plezierige momenten van mijn werk beschouwd (dit is echt veel leuker dan het klinkt!). Onder andere door deze momenten, durf ik met vrij grote zekerheid zeggen dat we een ijzersterk team vormden om allerhande fundamentele onderzoeksproblemen aan te pakken binnen de wachtlijntheorie.

Dieter, ook jouw deur stond altijd open voor vragen en discussies, waarvoor dank. Jij en Joris vulden elkaar goed aan op dat vlak. Naast jouw drive in het onderzoek naar nieuwe toepassingen voor de wachtlijntheorie, zet je je ook enorm in voor het onderwijs binnen de vakgroep. Iets wat ik enkel kan bewonderen.

Ook zou ik graag Bart Steyaert bedanken voor het delen van zijn kennis over staartbenaderingen in mijn beginperiode bij SMACS. Verder wil ik ook prof. Hans Vernaeve bedanken omdat hij zo vriendelijk was om de tijd te nemen om uitgebreid te antwoorden op enkele lastige vragen over complexe analyse.

De leden van de examencommissie wil ik uitdrukkelijk bedanken voor het evalueren van dit proefschrift.

Al konden ze me niet met inhoudelijke zaken helpen, toch had ik snel door hoe belangrijk ze zijn, de "stille" krachten achter TELIN: Davy, Philippe, Patrick en Sylvia. Davy en Philippe, bedankt om altijd zo snel mijn computer-gerelateerde problemen op te lossen. Davy, het was leuk om nog iemand te hebben die van dezelfde streek is en bovendien net zoals ik een fervente Pokémon trainer is. Patrick en Sylvia, het was een luxe om me nooit zorgen te moeten maken over onkostennota's en dergelijke. Patrick, extra bedankt dat jouw deur altijd open stond om het voorbije voetbalweekend te analyseren.

Bij SMACS werd ik steeds omringd door fantastische collega's. Daarom verdienen ook zij een woord van dank. Dankjewel Apoorv, Freek, Sarah, Sara, Hossein, Michiel, Caitlin, Kishor, Bart, Mustafa en Willem. De vakgroep telt nog vier andere sterke onderzoeksgroepen. Eén daarvan is DIGCOM, met wiens leden ik geregeld 's middags samen lunchte. Dankjewel Adriaan, Jelle, Johannes, Stef, Sander, Nele en Heidi.

Naast mijn promotoren en collega's, wil ik graag ook vrienden en (schoon)familie bedanken voor hun interesse in mijn onderzoek, maar eigenlijk nog meer voor de momenten die we samen doorbrachten die me even niet aan mijn werk deden denken. Mijn vrienden van de wiskunde, beter bekend als de partychat: ik hoop dat we samen nog veel spelletjesavonden doormaken en dat we eindelijk eens een quiz kunnen winnen! Mijn vrienden uit de Vlaamse Ardennen, beter bekend als de bende (al noem ik ze persoonlijk eigenlijk de spaghettibende): al woon ik nu wat verder, toch zal ik altijd met plezier blijven afkomen. De Ronde is van ons!

Tenslotte resten mij nog de drie belangrijkste personen in mijn leven om te bedanken. Mijn dankbaarheid voor hen gaat natuurlijk veel verder dan hun steun en interesse tijdens mijn onderzoek.

Mam & pap, dankjewel om er altijd te zijn voor mij. Jullie mogen trots zijn op jezelf. De vele kaarsjes en de gevleugelde woorden "Blijf kalm Jan" zullen nu een tijdje niet meer nodig zijn. Het is gerustellend om te weten dat ik altijd bij jullie terecht kan.

Hanne, alleen jij weet hoe gek ik echt ben (en op jou). Ik vind het nog altijd ongelooflijk hoe goed we elkaar begrijpen. Evenzeer is het moeilijk om te geloven hoe lang we al samen zijn. Eén ding weet ik zeker: we zijn wel degelijk OTP!

Arnaud Devos
Gent, April 2022

Contents

List of notations

Abbreviations

FCFS	first-come first-served
i.i.d.	independent and identically distributed
LST	Laplace-Stieltjes transform
PGF	probability generating function
pmf	probability mass function

Random variables

$a_{j,k}$	number of type-j arrivals during slot k ($j = 1, 2$)
r_k	number of available servers for type-1 customers during slot k
$u_{j,k}$	system content of type-j at the beginning of slot k
u_j	steady-state system content of type-j at the beginning of a random slot

Commonly used pmfs and PGFs

$a(i, j)$	joint pmf of the numbers of arrivals per slot
$A(z_1, z_2)$	joint PGF of the numbers of arrivals per slot
$A_l(z)$	PGF of the number of type-l arrivals per slot ($l = 1, 2$)
$p(i, j)$	joint pmf of the steady-state system contents of type-1 and type-2 at the beginning of a random slot
$p_l(n)$	pmf of the steady-state system content of type-l at the beginning of a random slot
$U_k(z_1, z_2)$	joint PGF of the system contents of type-1 and type-2 at the beginning of slot k
$U(z_1, z_2)$	joint PGF of the steady-state system contents of type-1 and type-2 at the beginning of a random slot
$U_l(z)$	PGF of the steady-state system content of type-l at the beginning of a random slot

System parameters

α	the probability that the server is available to type-1 customers during a random slot
λ_j	arrival rate of type-j customers

Mathematical operators

$(\cdot)^+$	$\max(\cdot, 0)$
$\mathbf{1}\{\cdot\}$	the identity operator
$\mathrm{corr}[\cdot,\cdot]$	the correlation coefficient between two random variables
$\mathrm{cov}[\cdot,\cdot]$	the covariance between two random variables
$\mathsf{E}[\cdot]$	the expectation operator
$\Pr[\cdot]$	the probability operator
$\mathrm{var}[\cdot]$	the variance of a random variable

The complex plane

i	the imaginary unit
$\mathrm{Re}[z]$	the real part of z
$\mathrm{Im}[z]$	the imaginary part of z
$\lvert z\rvert$	the absolute value of z
$\mathcal{R}_j$	the radius of convergence of $A_j(z)$

Partial derivatives

$$A^{(j)}(x,y) = \left.\frac{\partial A(z_1,z_2)}{\partial z_j}\right|_{z_1=x,z_2=y}$$

$$A^{(ij)}(x,y) = \left.\frac{\partial^2 A(z_1,z_2)}{\partial z_i \partial z_j}\right|_{z_1=x,z_2=y}$$

Samenvatting

In dit proefschrift bestuderen we een welbepaald wachtlijnmodel. Een wachtlijnmodel is een wiskundige beschrijving van een reëel wachtlijnsysteem. Onder een wachtlijnsysteem verstaan wij elk systeem waar entiteiten (die we aanduiden als de klanten) wachten op één of andere vorm van bediening. Een wachtlijnmodel wordt wiskundig beschreven aan de hand van een aantal (stochastische) elementen zijnde het aankomstproces, de wachtruimte en de werking van de bedieningsstations. Voor het merendeel van de wachtlijnmodellen is men geïnteresseerd in het berekenen van de stationaire distributie van het aantal klanten in het systeem. Eens deze distributie gevonden is, kunnen verschillende interessante prestatiematen van het systeem berekend worden.

Het wachtlijnmodel dat wij in dit proefschrift bestuderen, maakt onderscheid tussen twee verschillende types klanten, die wij aanduiden als klanten van type 1 en klanten van type 2. Voor beide types klanten is er een aparte wachtrij voorzien. Het is echter zo dat slechts één klant per keer bediend kan worden. Klanten van type 1 en type 2 kunnen dus niet tegelijkertijd bediend worden. Er is met name één bedieningsstation (*Engels*: server) dat verantwoordelijk is voor de bediening van beide wachtrijen. De volgorde waarin de verschillende types klanten bediend worden, kan op verscheidene manieren bepaald worden. De regels die deze volgorde beschrijven, worden samengevat door de 'scheduleringsdiscipline'. Het opdelen van klanten in verschillende types is noodzakelijk in wachtlijnsystemen waarbij klanten verschillende vereisten hebben. Een opdeling die bijvoorbeeld veel gebruikt wordt in computernetwerken is die van reële-tijdsverkeer (multimediatoepassingen zoals videobellen) ten opzichte van niet-reële-tijdsverkeer (zoals het versturen van een e-mail). Het eerste type verkeer (of m.a.w. het eerste type klanten) is gevoelig voor lange wachttijden, maar niet voor het verlies van 'klanten' (wanneer de wachtrij volzet is), terwijl het laatste type verkeer vereist dat er geen klanten worden verloren (en de wachttijden spelen typisch ook een minder grote rol). Het spreekt voor zich dat de scheduleringsdiscipline een belangrijke rol speelt in het voldoen aan de vereisten van beide types klanten.

In dit proefschrift bestuderen we grondig één specifieke scheduleringsdiscipline, namelijk waarbij de bediening tussen de twee types klanten op een willekeurige en afwisselende manier plaatsvindt. Telkens wanneer een nieuwe klant moet gekozen worden om te bedienen, wordt een muntstuk opgegooid waarvan de uitkomst kop is met kans α en munt met kans $1 - \alpha$. In het geval van kop,

kan een klant van type 1 bediend worden en in geval van munt kan een klant van type 2 bediend worden. In het geval dat er geen klanten aanwezig zijn van het gekozen type (bijvoorbeeld wanneer er kop wordt gegooid maar er zijn geen klanten van type 1), wordt er niemand bediend en wordt de munt gewoon opnieuw opgegooid. Het unieke bedieningsstation dat verantwoordelijk is voor beide wachtrijen, wordt op deze manier proportioneel verdeeld onder beide wachtrijen. Naargelang de waarde van α zal een wachtrij het unieke bedieningsstation meer of minder tot zijn beschikking hebben.

Voor dit wachtlijnmodel bestuderen we het aantal klanten van beide types *gezamenlijk*. Het is evident dat we dan kunnen spreken van een twee-dimensionaal wachtlijnmodel. Eveneens vanzelfsprekend is het feit dat het analyseren van twee-dimensionale wachtlijnmodellen beduidend lastiger is dan het analyseren van (klassieke) één-dimensionale wachtlijnmodellen. Toch is er al veel succesvol onderzoek verricht met betrekking tot algemene twee-dimensionale wachtlijnmodellen. Met behulp van probabiliteitsgenererende functies komt het onderzoeksprobleem steevast neer op het oplossen van een functionele vergelijking. Het moeilijke aspect aan het oplossen van deze vergelijking, is dat de vergelijking twee onbekende functies bevat (naast de te vinden functie). Dankzij eerder onderzoek is er aangetoond dat deze twee onbekende functies te bepalen zijn als de oplossing van een randwaardeprobleem voor holomorfe functies. Het numerieke werk dat gepaard gaat bij het berekenen van prestatiematen van het wachtlijnsysteem is echter zeer omslachtig met deze oplossingsmethode. Deze scriptie is bedoeld om het wachtlijnmodel, zoals beschreven in vorige alinea, beter te begrijpen. In het bijzonder besteden we veel aandacht aan de wiskundige moeilijkheid om dit model 'op te lossen' door verschillende oplossingsmethoden aan te reiken en speciale gevallen te beschouwen.

We bestuderen twee discrete toevalsveranderlijken in dit proefschrift, namelijk het aantal klanten in het systeem van type 1 en type 2, op een lukraak tijdstip 'in the long run'. Het doel is het berekenen van de stationaire gezamelijke massafunctie (*Engels*: probability mass function of pmf) van deze twee discrete toevalsveranderlijken. Om dit doel te bereiken, maken we gebruik van probabiliteitsgenererende functies (*Engels*: probability generating functions of PGFs). De complexiteit van de analyse blijkt sterk afhankelijk te zijn van de aard van het aankomstproces van de klanten. Voor een algemeen aankomstproces is het zelfs zo dat we niet in staat zijn om een exacte uitdrukking in gesloten vorm te vinden voor de gezamelijke PGF van het aantal klanten in het systeem van type 1 en type 2.

De twee doelstellingen van dit proefschrift luiden als volgt. Doelstelling 1: nagaan welke aankomstprocessen wél aanleiding geven tot een exacte oplossing in gesloten vorm; Doelstelling 2: bestuderen welke welgekende benaderingsmethoden voor één-dimensionale modellen succesvol kunnen worden uitgebreid en toegepast op ons twee-dimensionaal model in het geval van algemenere aankomstprocessen.

Hoofdstuk 2 is gericht op onze eerste doelstelling. In dit hoofdstuk beschou-

wen we drie specifieke aankomstprocessen. Met behulp van toepassingen uit de complexe analyse slagen we erin om de gezamelijke PGF van het aantal klanten in het systeem van type 1 en type 2 te vinden, voor elk van de drie specifieke gevallen. Aangezien de PGFs telkens rationale functies zijn, is het slechts een kleine moeite om ook de overeenkomstige pmfs terug te vinden. Cruciaal voor elk van deze drie analyses, is het concept van analytische uitbreiding van een complexwaardige functie. Deze techniek bleek overigens ook succesvol om asymptotische uitdrukkingen te bekomen voor de gezamelijke pmf van het aantal klanten in het systeem, voor algemenere aankomstprocessen (dit is het onderwerp van Hoofdstuk 3). De resultaten die we bekomen in Hoofdstuk 2 mogen dan wel enkel geldig zijn voor enkele zeer specifieke aankomstprocessen, toch hebben we inzicht gekregen in de structuur van de functionele vergelijking. Bovendien geven de bekomen uitdrukkingen voor de verschillende prestatiematen al een vrij goed inzicht in de impact van de scheduleringsdiscipline op het wachtlijnsysteem. We lichten één van deze inzichten nu toe. In het speciale geval dat de de klanten van type 1 en type 2 aankomen volgens twee onafhankelijke Bernoulli processen, dan is het zo dat de correlatie tussen het aantal klanten in het systeem van type 1 en van type 2 steeds negatief is, ongeacht de waarde van de systeemparameters (zijnde α en de gemiddelde aankomstintensiteiten). Dit is dan ook wat we intuïtief verwacht hadden. Wanneer één wachtrij zeer groot is, is dit te verklaren door twee zaken: ofwel zijn er veel aankomsten geweest de laatste tijd, ofwel heeft de wachtrij het bedieningsstation weinig tot zijn beschikking gehad (en kunnen er dus geen klanten vertrekken). In dit laatste geval is het dan zeer waarschijnlijk dat er weinig klanten in de andere wachtrij zijn, aangezien die het bedieningsstation veel tot hun beschikking hebben gehad.

Hoofdstuk 3, Hoofdstuk 4 en Hoofdstuk 5 focussen op de tweede doelstelling. Zoals eerder vermeld, worden in Hoofdstuk 3 asymptotische (dus geen exacte) uitdrukkingen gevonden voor de gezamelijke pmf van het aantal klanten in het systeem van type 1 en type 2, waar we ons niet langer beperken tot enkele specifieke aankomstprocessen zoals in Hoofdstuk 2. Met een asymptotische uitdrukking bedoelen we hier een uitdrukking voor de massafunctie wanneer het aantal klanten in één wachtrij naar oneindig gaat. Deze uitdrukkingen zijn zeer elegant en kunnen eenvoudig worden toegepast. De resultaten in dit hoofdstuk zijn weliswaar enkel geldig wanneer het aankomstproces voldoet aan een intrigerende voorwaarde. Deze voorwaarde is altijd voldaan wanneer de klanten van type 1 en type 2 aankomen volgens twee onafhankelijke aankomstprocessen. Vandaar dat deze resultaten algemener zijn dan deze die bekomen zijn in Hoofdstuk 2.

Hoewel de resultaten uit Hoofdstuk 3 elegante en efficiënte benaderingen opleveren, zijn deze niet accuraat genoeg voor het schatten van gezamelijke kansen dat er weinig klanten in het systeem aanwezig zijn. Niettegenstaande er reeds een uitgebreid gamma aan benaderingsmethoden voor twee-dimensionale wachtlijnmodellen bestaat, stellen we zelf een nieuwe benaderingsmethode voor in Hoofdstuk 4. Deze nieuwe benaderingsmethode combineert de resultaten uit

Hoofdstuk 3 met een interpolatiemethode. De combinatie van deze twee is origineel en verschillend in vergelijking met andere benaderingsmethoden. De resultaten bekomen met deze nieuwe methode werden vergeleken met simulaties. Er kan besloten worden dat de resultaten nauwkeurig zijn in het geval dat de bezettingsgraden (*Engels*: the load(s)) van de wachtrijen laag tot middelmatig zijn. Er wordt een verklaring gegeven waarom de resultaten onnauwkeurig zijn in het geval van hoge bezettingsgraden. Bovendien geven we ook suggesties om de benaderingsmethode nog te verbeteren in de toekomst.

Tenslotte, onderzoeken we in Hoofdstuk 5 nog een andere benaderingsmethode, namelijk een *heavy-traffic* benadering. Deze benaderingsmethode is zeer populair voor één-dimensionale wachtlijnmodellen, waar men typisch het aantal klanten in het systeem schaalt met de bezettingsgraad en vervolgens de bezettingsgraad naar 1 laat gaan. Dit noemt men een heavy-traffic limiet. In Hoofdstuk 5 veronderstellen we dat de klanten van type 1 en type 2 aankomen volgens twee gelijke en onafhankelijke aankomstprocessen en dat $\alpha = \frac{1}{2}$. Het nemen van de heavy-traffic limiet brengt ons nog steeds tot het oplossen van een functionele vergelijking. Echter, deze nieuwe functionele vergelijking is een pak eenvoudiger dan de oorspronkelijke en kan nu expliciet worden opgelost. Uit deze oplossing vinden we onder andere een uitdrukking in gesloten vorm voor de correlatiecoëfficiënt van het aantal klanten in het systeem van type 1 en type 2, wanneer de bezettingsgraad naar zijn kritische waarde 1 nadert. Aanvullend in Hoofdstuk 5 bestuderen we een ander, weliswaar gelijkaardig, wachtlijnmodel. Het verschil zit hem in het feit dat er nu altijd een klant wordt bediend zolang er klanten aanwezig zijn in het systeem. In de wachtlijntheorie zegt men dan dat het systeem *werkconserverend* (*Engels*: work-conserving) is. Herinner u dat in 'ons' wachtlijnmodel het zo kan zijn dat zich een mismatch kan voordoen wanneer het bedieningsstation beschikbaar is voor een wachtrij zonder klanten, terwijl er wel klanten aanwezig zijn in de andere wachtrij. We passen het model nu aan zodat deze mismatch zich niet meer kan voordoen. Met andere woorden, indien juist één wachtrij leeg is, dan zal het bedieningsstation altijd beschikbaar zijn voor de niet-lege wachtrij. Voor dit aangepast, werkconserverend, wachtlijnmodel berekenen we eveneens een heavy-traffic limiet. Beide modellen worden vervolgens met elkaar vergeleken, aan de hand van het gemiddeld totaal aantal klanten in het systeem en de correlatiecoëfficiënt tussen het aantal klanten in het systeem van type 1 en type 2. Het resultaat van deze vergelijkende studie is dat beide systemen significant verschillend zijn wanneer de bezettingsgraad kritiek is.

Summary

We present and analyze a specific queueing model. A queueing model is a mathematical representation of a system where entities (typically called customers) have to wait before receiving some kind of service. Such a model is defined by a number of stochastic processes, describing the arriving flow of customers, the waiting room and the service facilities. Ultimately, the target for many queueing models is to obtain the stationary (or steady-state) distribution of the number of customers in the system. From this distribution, numerical characteristics of interest can be calculated.

The queueing model studied in this dissertation has two different types of customers, each with their own dedicated queue, but there is only one single server facility. A maximum of one customer can be served at a time. At this point, a scheduling discipline is necessary. The scheduling discipline decides in which order the customers are handled. Customers of type 1 and type 2 thus compete for one and the same service. The differentiation of customers in different customer types can be necessary because they may have different Quality of Service requirements, or QoS, such as no loss, minimal delay, maximum bandwidth, etc. For example, in computer networks, typically two types of traffic are distinguished: real-time traffic (delay-sensitive, but loss-tolerant) such as videoconferencing; and non-real-time traffic (loss-sensitive, but delay-tolerant) such as file transferring. Ideally, both customer types strive for a maximum QoS. Providing QoS becomes an issue in case of limited resources, since in this case there is an obvious trade-off between the QoS of both customer types. It goes without saying that the choice of the scheduling discipline has a significant impact on the QoS in a network with limited resources.

In this dissertation, we consider the following scheduling discipline: at each service opportunity, a weighted coin is flipped so that with probability α a customer at queue 1 is served (if any) and with probability $1 - \alpha$ a customer at queue 2 is served (if any). If a queue happens to be empty at the moment that a service is allocated to that queue, no service occurs for a fixed amount of time. We call this discipline the randomly alternating service discipline. The single server facility is thus proportionally divided between the two customer types over time.

In the queueing model, we keep track of both the number of customers of type 1 and of type 2. For obvious reasons, our model is a so-called two-dimensional queueing model. The study of these models is known to be noto-

riously hard. Nevertheless, there has already been much research on (general) two-dimensional queueing models. Using the probability-generatin- function approach, a functional equation has to be solved. However, in order to solve this single functional equation, two unknown functions have to be determined. The standard state-of-the-art methodology exists of reducing it to a boundary-value problem for analytic functions (Dirichlet, Riemann, Riemann-Hilbert). However, in practice, cumbersome numerical work is necessary to obtain actual performance measures of the queueing model, such as the probability that there are more than a specified number of customers in the (total) system, the correlation coefficient between the numbers of customers of both types, etc. This dissertation is devoted to make several contributions towards a better understanding of the solution of the two-class queue with randomly alternating service.

The random variables analyzed in this dissertation are the numbers of customers of both types in the system in steady state, called the *system contents*. The objective is to obtain the joint distribution of these two random variables in steady state. The analysis makes extensive use of the theory of probability generating functions (PGFs). Obtaining easy-to-evaluate expressions for the probability generating function for this general queueing model proves to be unfeasible. The complexity of the analysis depends on the complexity of the arrival processes.

The two goals of this dissertation can be stated as follows. Goal 1: to investigate which arrival processes give rise to an exact closed-form expression of the joint PGF and the joint probability distribution of the system contents; Goal 2: to study and apply well-known one-dimensional approximation techniques to the two-dimensional problem at hand.

Chapter 2 is devoted to the first goal. In this chapter we consider three specific arrival processes. An expression for the joint PGF is obtained by using applications of complex analysis. Since all the obtained PGFs in this chapter are rational functions, the corresponding distributions are easily obtained. The notion of analytic continuation is not a standard one within queueing theory, but it is proven to be highly suitable for our case. Moreover, it can be applied under less severe restrictions of the arrival processes to obtain the *asymptotic* behavior of the joint distribution (instead of the exact expression). Despite the fact that the results in this chapter are only valid for specific arrival processes, these give valuable insights in the analysis of the functional equation. Moreover, the simple expressions allow to show the impact of the scheduling discipline on the performance of the queueing model. We point out one of such finding. In the case that customers of type 1 and type 2 arrive according to two independent Bernoulli processes, it turns out that the system contents are negatively correlated (at least for the arrival distribution under consideration). This is in accordance with our intuition. If the number of customers in the first queue is exceptionally large, then either there have been a lot of arrivals lately to the first queue, or the first queue is not served often lately. In the latter

case, it is likely that the number of customers in the second queue is small.

Chapter 3, Chapter 4 and Chapter 5 are devoted to the second goal. As mentioned before, analytic continuation can be applied to obtain the *asymptotic* behavior of the joint distribution of the system contents (instead of the exact expression), without having to restrict ourselves to specific arrival processes. Such an asymptotic analysis is the subject of Chapter 3. In this chapter, we introduce an intriguing condition for the arrival process in order to compute the *tail asymptotics* of the joint distribution of the system contents. Broadly speaking, with tail asymptotics we mean an expression for the joint distribution of the system contents, when the number of customers in one queue is considered to be large. The results of this chapter are important, since the numerical work to calculate these obtained expressions is negligible (absolutely and relatively as compared to the boundary-value approach).

While the results of Chapter 3 serve as an elegant and efficient approximation technique, it is obviously inaccurate to estimate the (joint) probabilities of *small* system contents. Although a myriad of approximation schemes exist for two-dimensional queueing models, we proposed a novel approximation method which is presented in Chapter 4. Interpolation methods and tail asymptotic results are combined to approximate the complete joint distribution of the system contents. The combination of these two concepts is the novelty and difference of our approach in comparison with previous studies. The results of this approximation method are compared with simulation results. In case of low and medium load, accurate results are obtained. We explain why the results are inaccurate in case of high loads. Moreover, we also suggest how to improve these inaccuracies in the future.

In Chapter 5, we determine a heavy-traffic limit. This is a renowned technique and is widely applied to one-dimensional queueing models. For multi-dimensional (or two-dimensional) models, this turns out to be more difficult yet again in comparison with the one-dimensional case. In Chapter 5, we assume that customers of type 1 and type 2 arrive according to two equal and independent arrival processes and $\alpha = \frac{1}{2}$. We linearly scale the steady-state system contents and derive a functional equation for the corresponding joint Laplace-Stieltjes transform. After taking the heavy-traffic limit, we show that the functional equation can be solved explicitly by means of the boundary-value approach. Several mixed moments of interest can be computed from the joint Laplace-Stieltjes transform. The most interesting one is the correlation coefficient. Because the correlation coefficient is invariant under linear transformations of the random variables, we have obtained an explicit expression of the correlation coefficient between the system contents when the queueing system is brought to the border of instability. In addition, we compare the two-class queue with randomly alternating service with its *work-conserving* variant. A work-conserving scheduling discipline is one that always serves a customer when there is a customer in the system. The randomly alternating service discipline is not work-conserving because there can be customers from one queue

waiting while the server is allocated to an 'empty' queue. It has to be said that most scheduling strategies examined in the literature are assumed to be work-conserving. The most prominent reason to assume a non-work-conserving scheduler is to keep the frequency of switching between the two types at a predetermined level (this is an advantage when there are costs involved with switching), while still guaranteeing that no customer types suffer from starvation. In the work-conserving variant of the randomly alternating scheduler, it is assumed that when only one of both queues is non-empty and the other is empty, the non-empty queue is served. For both schedulers, we compared the mean total system content and the correlation coefficient between the system contents, in heavy-traffic. The result of this study is that both schedulers are significantly different from each other.

1

Introduction

1.1 What is queueing theory?

Queueing theory is literally the scientific study of queueing phenomena. Since the input and output variables of a queue are usually of a random, non-deterministic nature, queueing theory mainly uses the apparatus of probability theory. Queueing theory is considered to be one of the oldest, and also most notable and prominent, subareas of the field of (applied) probability theory [1, Ch. III]. Additionally, it should be mentioned that queueing theory dates back to the pioneering work of the Danish mathematican Agner Krarup Erlang (1878-1929) who used -what we now call- queueing theory to establish how many operators would be necessary in a telephone exchange so as to avoid overly long waiting times. Agner Krarup Erlang is widely recognized as the founder of queueing theory.

When thinking about "a queue", it is natural to consider actual humans that are standing in line, patiently waiting to receive some kind of service from a cashier. However, there are many other kinds of queueing situations from daily life. To name a few: we wait patiently to be answered by technical support when we call a phone service provider; clear the security check at an airport; experiencing delays while browsing the world wide web; and so on. Queueing models are useful to performance modeling and analysis of (tele)communication networks, transportation systems, manufacturing systems, and in other fields that involve scheduling and logistics [2, Ch. 23]. Without going into great detail, we want to draw attention to the application of queues in communication systems since the evolution of queueing theory is intimately tied to communication systems. In communication systems, multiple information units are sent over shared links. Queues are provided to store the information units temporarily during time periods that more information arrives than can be simultaneously transmitted. Within the performance evaluation community, typically other terminology is used than the classic queueing terminology. The most prominant examples are *buffers* instead of queues and *packets* instead of customers. According to [3,

Ch. 1], the goals of performance analysis are twofold. The first one is *predicting* the system performance. For example, one wants to estimate the probability that the queue length exceeds a given threshold. While prediction is important and useful, an even more important goal for a system engineer is finding a superior design to improve the system performance. Mathematical queueing models allow for getting insight into the dynamics of a queueing process prior to its costly implementation. Or, put differently, queueing models are to a system engineer what architectural models are to architects.

We emphasize that this dissertation is of a fundamental nature. A particular mathematical queueing model can be used for multiple applications, but a fundamental queueing theorist is in the first place interested in capturing the essential aspects of the queueing problem at hand, regardless of the application. Throughout this dissertation, the emphasis is put on the mathematical analysis of a particular queueing model, rather than on the possible applications as described above.

To conclude this introductory section on queueing theory, we want to remark that there are two main analytical approaches to analyze queueing models, namely the *transform method* and the *matrix-analytic method*. The latter translates the queueing problem at hand to a linear system with an infinite number of unknowns -usually the stationary probabilities of the system contents- and an infinite number of linear equations in these unknowns. By exploiting the structural properties of the (infinite) coefficient matrix, computationally efficient algorithms are developed. We refer to [4–6] for further details. The transform method makes extensive use of the theory of probability generating functions and Laplace-Stieltjes transforms. Which analytical approach out of these two to choose is often a matter of taste. Broadly speaking, researchers who prefer the matrix-analytic method are more interested in stable algorithms and linear algebra, while researchers with a preference for the transform method are likely more interested in closed-form expressions and in mathematical analysis (in which we mean that branch of mathematics studying functions, limits, derivatives, etc.). Throughout this dissertation we will use the transform approach. Good walking shoes, i.e. a solid introduction to probability theory and complex analysis, are all the equipment that is needed to walk trough this dissertation.

1.2 From probability theory to complex analysis

Throughout this dissertation we heavily rely on the use of generating functions -in particular probability generating functions and Laplace-Stieltjes transforms- to analyze random variables of interest. In fact, we even dare to say that generating functions are the central object of this dissertation. Therefore, in this section we list some of the main properties of probability generating functions and Laplace-Stieltjes transforms. In particular, we already want to draw at-

tention to the fact that a crucial part of the analysis of this dissertation is to interpret generating functions as functions of a complex variable. Most part of this section is based on the books [1, 7–11].

The probability generating function (PGF) of a discrete random variable X, with values in $\mathbb{Z}_{\geq 0}$, is by definition

$$X(z) \triangleq \mathsf{E}[z^X] \tag{1.1}$$

$$= \sum_{k=0}^{\infty} \Pr[X = k] z^k \,, \tag{1.2}$$

where $\mathsf{E}[\cdot]$ denotes the expectation operator and $\Pr[\cdot]$ the probability measure. Note that speaking of the PGF of a random variable is abuse of language. More correctly would be to say that $X(\cdot)$ is the generating function of the probability mass function $\Pr[X = \cdot\,]$. However, in this dissertation we will always speak about the PGF of a random variable, referring to the complex-valued function with power series representation (1.2).

PGFs have many useful properties. Since all probabilities must sum up to one, we have that

$$X(1) = 1 \,. \tag{1.3}$$

The moment generating property of PGFs allows one to obtain factorial moments of the random variable, i.e.

$$\mathsf{E}[X(X-1)\dots(X-k+1)] = \frac{d^k}{dz^k} X(z)\bigg|_{z=1} \,. \tag{1.4}$$

For example for $k = 1$ and $k = 2$, we get

$$\mathsf{E}[X] = \frac{dX(z)}{dz}\bigg|_{z=1} \,, \quad \mathsf{E}[X(X-1)] = \frac{d^2X(z)}{dz^2}\bigg|_{z=1} \,. \tag{1.5}$$

The variance of X is thus found as

$$\begin{aligned} \operatorname{var}[X] &= \mathsf{E}[X^2] - \mathsf{E}[X]^2 \\ &= \frac{d^2X(z)}{dz^2}\bigg|_{z=1} + \frac{dX(z)}{dz}\bigg|_{z=1} - \left(\frac{dX(z)}{dz}\bigg|_{z=1}\right)^2 \,. \end{aligned} \tag{1.6}$$

Equations (1.5) and (1.6) provide the simplest means to calculate $\mathsf{E}[X]$ and $\operatorname{var}[X]$, provided that an explicit expression for the PGF $X(z)$ is available. Higher moments of X can of course also be obtained by taking the appropriate derivatives and evaluating at $z = 1$.

We now come to the complex analysis part. According to (1.2), $X(z)$ is a power series in z with coefficients $\Pr[X = k]$. Let $\mathcal{R}$ be the radius of convergence of $X(z)$. For a function $X(z)$ that has a Taylor series with non-negative coefficients (which is the case for PGFs), Pringsheim's theorem [11, Th. IV.6] states

that $\mathcal{R}$ is a singularity of the function $X(z)$. Furthermore, $\mathcal{R} \geq 1$ since the power series (1.2) converges absolutely for $|z| \leq 1$. Singularities of $X(z)$ which lie on the boundary of the disc of convergence, are called *dominant singularities*. Or in other words, dominant singularities are singularities with smallest norm. It has to be said that in practice, most of the time there is a unique dominant singularity (which is then necessarily equal to the radius of convergence $\mathcal{R}$).

Let us now introduce the concept of analytic continuation. A function $X(z)$ which is given originally in the form of a power series with a finite radius of convergence can be investigated beyond the circumference of the circle of convergence by a procedure termed *analytic continuation*. A more formal definition of this concept, based on [7, Ch. III], is given by:

Definition 1.1 (Analytic continuation)**.** *Let $X(z)$ be an analytic function defined over Ω. If there exists an analytic function $X^\star(z)$ defined over some open set $\Omega^\star$, with $\Omega \cap \Omega^\star \neq \emptyset$, and such that $X^\star(z) = X(z)$ in $\Omega \cap \Omega^\star$, one says that X is analytically continuable in $\Omega^\star$ and that $X^\star$ is the analytic continuation of X at $\Omega^\star$.*

A very useful theorem in the concept of analytic continuation is the following theorem (based on [10, Th. 3.2.6]).

Theorem 1.1. *Let $X(z)$, $Y(z)$ be analytic in the open domain Ω. If $X(z) = Y(z)$ in some subportion $\Omega' \subset \Omega$, then $X(z) = Y(z)$ everywhere in Ω. In particular it is sufficient that $X(z)$ and $Y(z)$ coincide on a curve interior to Ω.*

Using Theorem 1.1, it can be proven that analytic continuation is unique. Analytic continuation is often established using a functional equation. To illustrate this, we cannot think of a better example than that of the gamma function, defined by $\Gamma(z) = \int_0^{+\infty} t^{z-1}e^{-t}dt$. One can easily show that this integral converges for $\mathrm{Re}(z) > 0$. Consequently, $\Gamma(z)$ is an analytic function on the right half-plane $\mathrm{Re}(z) > 0$. The identity $\Gamma(z) = \frac{\Gamma(z+1)}{z}$, valid for real $z > 0$ (obtained via partial integration), remains valid for complex z for which $\mathrm{Re}(z) > 0$ by Theorem 1.1. We can use this functional equation to analytically continue $\Gamma(z)$ as follows. The function $g(z) := \frac{\Gamma(z+1)}{z}$ is an analytic function on $\{z \in \mathbb{C} \ : \ \mathrm{Re}(z) > -1\} \setminus \{0\}$ which agrees with Γ on the right half-plane $\mathrm{Re}(z) > 0$. If we keep denoting this extended function with Γ, then $g(z)$ is well defined and analytic in $\{z \in \mathbb{C} \ : \ \mathrm{Re}(z) > -2\} \setminus \{0, -1\}$, etc. until we get a unique analytic continued function on $\mathbb{C} \setminus \{0, -1, -2, \ldots\}$.

Generating functions can also be defined for a pair of random variables X and Y, with values in $\mathbb{Z}_{\geq 0}$. The PGF of X and Y is given by

$$P(z_1, z_2) \triangleq \mathsf{E}[z_1^X z_2^Y] \tag{1.7}$$

$$= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \Pr[X = k, Y = l] z_1^k z_2^l \, . \tag{1.8}$$

Such a generating function will be called the joint PGF of X and Y. In this context, we will call (1.1) the *marginal* PGF of X. Cross-moments between X and Y can be computed from $P(z_1, z_2)$. For instance, the covariance between X and Y is given by

$$\mathrm{cov}[X,Y] = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y] \tag{1.9}$$

$$= \left.\frac{\partial^2 P(z_1,z_2)}{\partial z_1 \partial z_2}\right|_{z_1=1,z_2=1} - \left(\left.\frac{dX(z)}{dz}\right|_{z=1}\right)\left(\left.\frac{dY(z)}{dz}\right|_{z=1}\right) , \tag{1.10}$$

with $X(z)$ and $Y(z)$ the marginal PGFs of X and Y, respectively.

We point out three properties that are evident from definition (1.7). First, the marginal PGFs of X and Y can be found from the joint PGF as $X(z) = P(z,1)$ and $Y(z) = P(1,z)$, respectively. Secondly, the marginal PGF of the sum $X+Y$ is obtained as $P(z,z)$. Finally, the variables X and Y are independent if and only if $P(z_1,z_2) = X(z_1)Y(z_2)$.

We now turn our attention to Laplace-Stieltjes transforms. The Laplace-Stieltjes transform (LST) is a widely used integral transform with many applications in applied mathematics, and in particular in applied probability theory to study continuous random variables. The LST of a random variable X with values in $\mathbb{R}_{\geq 0}$ is by definition

$$X(s) \triangleq \mathsf{E}[e^{-sX}] \tag{1.11}$$

$$= \int_0^\infty e^{-st} dF_X(t) , \tag{1.12}$$

with F_X the cumulative distribution function of the random variable X. If X is a continuous random variable with density function f_X, then the LST (1.12) corresponds to the Laplace transform of f_X. Since $\lim_{x\to+\infty} F_X(x) = 1$ and $F_X(0) = 0$, we then have the normalization condition

$$X(0) = 1 . \tag{1.13}$$

Furthermore, the LST exhibits the moment generating property as follows

$$\mathsf{E}[X^n] = (-1)^n \left.\frac{dX(s)}{ds^n}\right|_{s=0} . \tag{1.14}$$

It is known from Laplace transform theory that if the LST converges at $s = s_0$, then it automatically converges for all s with $\mathrm{Re}[s] > \mathrm{Re}[s_0]$. Since $X(0) = 1$, the region of convergence is (at least) the half-plane $\mathrm{Re}[s] > 0$. In particular, it is analytic in this region. The concept of dominant singularities applies for LSTs as well. The dominant singularities of an LST will always be located on a line $\mathrm{Re}[s] = -a$, with $a \geq 0$. Furthermore, the unique real value a on this line, will always be a singularity of the LST.

LSTs can also be defined for a pair of continuous random variables X and Y with joint density function f_{XY} as follows

$$F(s_1, s_2) = \mathsf{E}[e^{-s_1 X} e^{-s_2 Y}] \tag{1.15}$$

$$= \int_0^\infty \int_0^\infty e^{-s_1 x} e^{-s_2 y} f_{XY}(x, y) dx dy \; . \tag{1.16}$$

Cross-moments between X and Y can be computed from $F(s_1, s_2)$. For instance, the covariance between X and Y is given by

$$\mathrm{cov}[X, Y] = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y] \tag{1.17}$$

$$= \left. \frac{\partial^2 F(z_1, z_2)}{\partial s_1 \partial s_2} \right|_{s_1=0, s_2=0} - \left(\left. \frac{dX(s)}{ds} \right|_{s=0} \right) \left(\left. \frac{dY(s)}{ds} \right|_{s=0} \right) \; , \tag{1.18}$$

Furthermore, we also point out three properties that are evident from definition (1.15). First, the marginal LSTs of X and Y can be found from the joint LST as $X(s) = F(s, 0)$ and $Y(s) = F(0, s)$. Secondly, the marginal LST of the sum $X + Y$ is obtained as $F(s, s)$. Finally, the variables X and Y are independent if and only if $F(s_1, s_2) = X(s_1)Y(s_2)$.

For the remainder of this section, we list some theorems from complex analysis that are frequently used throughout this dissertation. The first one is Rouché's theorem [10, Th. 4.4.2]. We remark that the use of Rouché's theorem is quite common in the analysis of queueing models via the transform method.

Theorem 1.2 (Rouché)**.** *Let f and g be two analytic functions inside and on a closed contour C in the complex plane such that $|g(z)| < |f(z)|$ for all z on C. Then the functions f and $f + g$ have the same number of zeros inside C.*

Another classic theorem from complex analysis that can be found in many textbooks is Liouville's theorem. However, application of this theorem is less common within queueing theory in comparison with Rouché's theorem. We first state Liouville's original theorem (see for example [9, Page 122]).

Theorem 1.3 (Liouville)**.** *A function which is analytic and bounded in the whole complex plane must reduce to a constant.*

In fact, we will also use the following extended version of Liouville's theorem in this dissertation.

Theorem 1.4 (Extended Liouville)**.** *Suppose f is a function which is analytic in the whole complex plane and $|f(z)| < M|z|^m$ for sufficiently large $|z|$, then f must reduce to a polynomial of degree at most m.*

1.3 Two-class queueing systems and scheduling

Two-class queueing systems are queueing systems where two types of customers need (and compete for) the same service. The concept of having two different

customer types can be based on any kind of binary classification variable. Some concrete examples of such a binary variable are: male or female (in case of human customers), real-time traffic or non-real-time traffic (in case of computer networks), premium or regular service, urgent or non-urgent service, etc. A two-class queueing system can be pictured as maintaining two separate subqueues for the two customer types. A *scheduling discipline* of the server then decides in which order the customer types have access to the server, as shown in Figure 1.1. Once the server has chosen a queue, it can take a customer from that queue on a first-come-first-served (FCFS) basis. We say then that the server takes the *oldest* customer from that queue, this is the customer with the longest residing time in that queue.

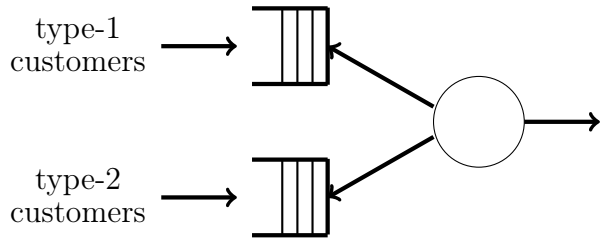


Figure 1.1: When the server becomes free, the next customer to take place in the server is decided by the scheduling discipline.

Different scheduling strategies give rise to different queueing behavior. We start with a list of some of the main types of scheduling strategies in case of a two-class queueing system.

Global First Come First Served The customers are served on a FCFS basis, regardless of their type. The scheduling rule is thus that at each service opportunity, the oldest customer in the system is served [12, 13]. For the analysis of a two-class queueing system with global FCFS and a discrete-time parameter we refer to [14]. More recently, in [15], the discrete-time two-class queueing system with global FCFS and batch service is analyzed.

Fixed Priority A customer of the highest available priority type is served when a new service opportunity occurs. Therefore, customers of lower priority classes have no (or only a limited) influence on the number of customers of higher priority in the system (see e.g. [16, 17] and references therein for the treatment of queues with fixed-priority scheduling with two customer types).

Round Robin When both queues are non-empty, the next customer type to be served is the opposite type as the previous served customer type. When only one of both queues is non-empty and the other is empty, the non-empty

queue is served. The server alternately serves thus one type-1 customer and one type-2 customer (if any). This scheduling discipline was introduced to give each queue a guaranteed share of the available server capacity.

Random When both queues are non-empty, the server serves a customer according to the outcome of a single Bernoulli trial (with fixed probabilities). When only one of both queues is non-empty and the other is empty, the non-empty queue is served. In discrete-time queueing theory, this scheduling discipline can be seen as a discrete-time probabilistic version of Generalized Processor Sharing (GPS) (see e.g. [18, 19] for the discrete-time versions and see [20] for the original version of GPS). Just as with Round Robin, the random scheduling discipline gives each queue a guaranteed share of the available server capacity.

It is worth noting that our definition of two-class queueing systems can be more or less seen as a subclass of *polling systems*. In this type of systems, there are M servers, which are shared by N customer types. The order in which the servers serve these queues, i.e., the rules by which the next queue to be served (after the service of queue i has been completed) is determined, is called the *polling order*, while the rule by which the server decides to stop servicing a given queue is called the *queue service discipline* [21–25]. We emphasize that there exist many queue service disciplines and polling orders, see e.g. [23]. The most usual polling orders are cyclic and random. The most popular queue service disciplines in case of a single-server are the exhaustive, gated, number-limited and time-limited discipline. In the **exhaustive** service discipline, the server serves customers until the queue is emptied. See for example [26], where the time when the server switches queues is non-zero. In [27], the exhaustive scheduling policy is analyzed with more than two customer types. A **gated** discipline places a (fictitious) gate behind the customers present in the queue when the server arrives at this queue, and only the customers in front of the gate will be served before the server goes to the next queue. In contrast to the exhaustive discipline, newly arriving customers are not served during a visit of the server. In a polling system with a **number-limited** discipline, the server serves k customers at a queue (if any). The number k can be either a deterministic or a random variable. The hard-to-analyze two-queue model with 1-limited service is studied in [28, 29]. Finally, in the **time-limited** discipline, the server serves customers at the current queue until a time limit T expires. This time limit T can again be either a deterministic or a random variable. In both the number- and time- limited disciplines, the server also typically leaves the current queue when it becomes empty. For further references on time-limited polling systems, we refer to the works of [30–34]. For $M = 1$, $N = 2$ and a 1-limited service discipline our definition of two-class queueing systems can be mapped to that of polling systems (if we allow for the polling order that the same queue can be served again). For example, the 1-limited service discipline with a cyclic polling order is equivalent to round-robin scheduling and the 1-limited service discipline with a random polling order is equivalent

to random scheduling.

The most common scheduling strategies depend on the number of customers present in the system. For example, with fixed-priority scheduling, the server picks a low-priority customer as soon as the number of high-priority customers is zero. Thus the scheduling discipline here depends on the number of high-priority customers. Likewise, with round robin and random scheduling, the scheduling discipline depends on both the number of customer types since the server "switches" when a queue becomes empty. Most of the time, the scheduling strategies depend on the number of customers present in the system in order to keep the system *work-conserving*. A work-conserving scheduling discipline is one that always serves a customer when there is a customer in the system. Obviously, all the aforementioned scheduling strategies are work-conserving. To the best of our knowlegde, one of the first scheduling strategies that does not possess the work-conserving property and that is analyzed in the context of server-sharing models is found in the Fixed Time Loop System (FTLS) [35]. In the FTLS, a multi-class queueing system is considered where the server visits the queues for a fixed amount of time in a deterministic, cyclic order. The FTLS discipline is motivated by the fact that it is applicaple to systems where the switching rule is set by the system manager (the server) and not by the customers. This enables to keep the frequency of switching between the queues at a predetermined level (this is an advantage when there are costs involved with moving the server) and no queues suffer from starvation (which can happen with global-FCFS and fixed-priority scheduling). Without going into a detailed description, we refer to [36–39] for more recent scheduling strategies that do not possess the work-conserving property.

In this dissertation, we analyze a simple, conceptual, discrete-time two-class system where the scheduling discipline is like the random scheduling discipline, except that now the server is *available* to a certain customer type according to certain probabilities, *regardless* of the number of customers in the queue. We emphasize that there can be customers from one type waiting while the server is dedicated to an empty type. Hence, this randomly alternating service discipline is part of the class of non-work-conserving ones discussed in the previous paragraph.

1.4 Queueing model

In this section, we define the basic stochastic processes of the queueing model under investigation in this dissertation. A mathematical description of the arriving flow of customers, the waiting room and the service facilities stand as a basis of any queueing model. On top of that, the time parameter (discrete or continuous), can also be considered as a major component of a queueing model. Firstly, we specify the time parameter that is used in this dissertation and provide some background information. Secondly, the arrival process is

defined. Thirdly, we discuss the assumption with respect to the waiting area and finally the service process is characterized. The scheduling discipline is logically incorporated in the service process, but we will treat this separately since this is the main feature of the queueing model.

1.4.1 Time parameter

We consider a *discrete-time* queueing model. That is, the time axis is divided into fixed-length intervals referred to as (time) slots. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e. can only start and end at) slot boundaries. This means that an arriving customer cannot enter the server during its arrival slot, even when the server is empty when the customer arrives. In the literature, this is sometimes referred to as the late-arrival system with delayed access [16].

1.4.2 Arrival process

The arrival process of a queueing model characterizes how new customers enter the queueing system. In most queueing models, the arrival process is characterized by the inter-arrival times, which are defined as the time intervals between two consecutive (batch) arrivals. If customers arrive in batches, i.e. multiple customers arrive in the same slot, then the batch sizes must also be specified, otherwise it is assumed that only single arrivals occur. In this dissertation, we assume that the inter-arrival times constitute a series of independent and identically distributed (i.i.d.) random variables with common geometric distribution and the batch sizes constitute a series of i.i.d. discrete random variables (which can have any generic distribution). Because the geometric distribution is memoryless, this assumption is equivalent to that of assuming that the numbers of arrivals in consecutive time slots constitute a sequence of i.i.d. non-negative discrete random variables (which can have any generic distribution). Therefore, we may alternatively use this characterization as well. It is this latter characterization that we will use in the following, but we think it is useful for the reader to indicate the equivalency. Note that in both characterizations, the exact moment at which customers arrive within slots is not specified. However, this is of no importance since in discrete-time queueing models, system changes are only observed at slot boundaries. We refer to [40] for a more elaborate discussion on this topic. Finally, the i.i.d. nature of the arrival process is a typical assumption in discrete-time queueing theory, but we emphasize that this does not have to be the case. For examples how to model arrivals that are correlated during consecutive slots, we refer to [41–44] and references therein.

We will now define the arrival process of the queueing model studied in this dissertation and fix notations. Two types of customers, named type-1 and type-2, enter the system. We denote the numbers of arrivals of type-j during a slot k by $a_{j,k}$ $(j = 1, 2)$. Both types of customer arrivals are assumed to

be i.i.d. from slot to slot and are characterized by a common joint probability mass function (pmf) $a(i,j)$ and common joint probability generating function (PGF) $A(z_1,z_2)$ respectively. More specifically,

$$a(i,j) \triangleq \Pr[a_{1,k}=i, a_{2,k}=j], \quad i,j \geq 0, \tag{1.19}$$

and

$$A(z_1,z_2) \triangleq \mathsf{E}[z_1^{a_{1,k}} z_2^{a_{2,k}}] \tag{1.20}$$

$$= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} a(i,j) z_1^i z_2^j \,. \tag{1.21}$$

We denote the marginal PGFs of the number of type-1 and type-2 arrivals per slot by

$$A_1(z) \triangleq \mathsf{E}[z^{a_{1,k}}] \tag{1.22}$$

$$= A(z,1) \tag{1.23}$$

and

$$A_2(z) \triangleq \mathsf{E}[z^{a_{2,k}}] \tag{1.24}$$

$$= A(1,z) \tag{1.25}$$

respectively. The mean numbers of arrivals of type-1 and type-2 per slot, in the sequel referred to as the *(mean) arrival rates* of type 1 and type 2 respectively are given by

$$\lambda_j \triangleq A_j'(1) \quad j=1,2. \tag{1.26}$$

1.4.3 Queue(s) and queue capacity

Arriving customers are stored in one or multiple queues. For mathematical convenience, we assume that all customers are able to enter the system. Equivalent to this, it means that we assume that the queue(s) have infinite storage capacity. Because of this assumption, it becomes irrelevant for the mathematical analysis whether we assume that both customer types are stored in a common queue or if we assume two separate queues. In the remainder, we will always say there are two separate queues since this seems more logical to bear in mind the general picture of Figure 1.1. For further references, we refer to the queue of type-1 and type-2 customers as queue-1 and queue-2, respectively. Summarized, we thus assume that all arriving customers can enter their dedicated queue and will eventually be served and leave the system, if the system is stable.

1.4.4 Servers and service process

We assume that the service area of the queueing system consists of a single server, who is responsible for the service of both customer types. Consequently,

this means that at most one customer can be served during a slot. Further, we assume that the service of each customer requires exactly only one slot, regardless of whether the customer is of type-1 or type-2.

Although logically incorporated in the service facility, we will put a special emphasis on the characterization of the scheduling discipline, i.e. the rule determining the order in which customers are served (type-1 or type-2), in a separate section below.

1.4.5 The scheduling discipline: randomly alternating

At the beginning of every time slot, the single server randomly selects either queue-1 or queue-2 to serve. This selection occurs independently of the system state and is modeled by a single parameter α ($0 < \alpha < 1$), that is defined as

$$\alpha \triangleq \Pr[\text{server is available to type-1 customers during a slot}] .$$

This directly means that the server is available to type-2 customers during a slot with probability $1-\alpha$. Moreover, it is assumed that the state of the server (available to either queue-1 or queue-2) during a certain slot is independent of the state of the server during previous slots, and also of the other random variables present in the model.

More concretely, let r_k denote the number of available servers (0 or 1) for queue-1 during slot k. Consequently, the random variable $1 - r_k$ denotes the number of available servers for queue-2. By assumption, the sequence $\{r_k\}_{k\in\mathbb{N}}$ constitutes a sequence of i.i.d. Bernoulli random variables with probability mass function

$$\begin{aligned} \Pr[r_k = 0] &= 1 - \alpha \\ \Pr[r_k = 1] &= \alpha , \end{aligned}$$

and generating function

$$1 - \alpha + \alpha z .$$

It can be easily shown that the parameter α has the following physical meaning: the fraction of time that the server is available to queue-1.

Finally, we want to draw attention to the following non-work-conserving property of the scheduling discipline. Implicitly, we have assumed that when the server is available to an empty queue, no service occurs in that slot, even when the other queue is non-empty. Hence, the scheduling discipline is not work-conserving.

1.5 Analysis of the single-queue model

The marginal distributions of the numbers of customers present in both queues in this model can be obtained in a relatively easy way, since from the perspective

of one queue, this is a single-server single-class queueing system with server interruptions. In such a model, it is assumed that the server is subjected to random interruptions such that no service can occur during these slots. In our case, the interruption process is modeled by means of a sequence of i.i.d. Bernoulli random variables. This queueing model is in the classic textbook [40, Ch. 3.2] referred to as a *Bernoulli model*.

It is worth noting that the single-queue Bernoulli model with single-slot service times can also be seen as a single-queue model with geometric service times. This can be understood as follows. Without loss of generality, let us focus on type-1 customers. If there is at least one type-1 customer to serve, then one type-1 customer is served with probability α and no type-1 customers are served with probability $1-\alpha$, independently of the previous time slots. Hence, the *effective* service time of a type-1 customer is geometrically distributed with parameter $1-\alpha$, i.e.

$$\Pr[\text{service-time of a type-1 customer} = n \text{ slots}] = \alpha(1-\alpha)^{n-1}, n = 1, 2, \ldots$$

In the remainder of this section, we show the analysis of the single-queue Bernoulli model. We feel it is useful to do so, since it permits us to show some elementary techniques of discrete-time queueing theory using transforms. Moreover, we hope to convince the reader of the difference in difficulty of analyzing the queues separately compared to analyzing both together. To that end, let us define

$$u_{1,k} \triangleq \text{the number of type-1 customers at the beginning of slot } k \, .$$

Furthermore, the PGF of $u_{1,k}$ is denoted by $U_{1,k}(z)$, i.e.

$$U_{1,k}(z) \triangleq \mathsf{E}[z^{u_{1,k}}] \, . \tag{1.27}$$

The evolution of the number of type-1 customers is described by the following system equation (cf. Fig. 1.2):

$$u_{1,k+1} = (u_{1,k} - r_k)^+ + a_{1,k} \, , \tag{1.28}$$

where $(\cdot)^+$ denotes the maximum of the argument and 0.

Using generating functions, equation (1.28) yields

$$\begin{aligned} U_{1,k+1}(z) &= \mathsf{E}[z^{u_{1,k+1}}] \\ &= \mathsf{E}[z^{(u_{1,k}-r_k)^+ + a_{1,k}}] \\ &= \mathsf{E}[z^{(u_{1,k}-r_k)^+}]\mathsf{E}[z^{a_{1,k}}] \, , \end{aligned}$$

where we used that the random variable $a_{1,k}$ is statistically independent with respect to $u_{1,k}$ and r_k. The second factor $\mathsf{E}[z^{a_{1,k}}]$ in the equation above is

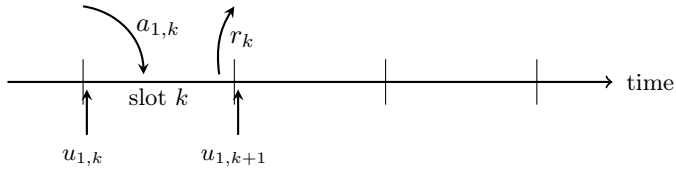


Figure 1.2: Time axis to illustrate the system equations.

nothing else than the pgf $A_1(z)$. The first factor can be calculated, using that $u_{1,k}$ and r_k are independent:

$$\begin{aligned}\mathsf{E}[z^{(u_{1,k}-r_k)^+}] &= \mathsf{E}[z^{(u_{1,k})^+}](1-\alpha) + \mathsf{E}[z^{(u_{1,k}-1)^+}]\alpha \\ &= U_{1,k}(z)(1-\alpha) + \frac{U_{1,k}(z) + (z-1)U_{1,k}(0)}{z}\alpha \; .\end{aligned}$$

We thus obtained the following relation between $U_{1,k+1}(z)$, $U_{1,k}(z)$ and $U_{1,k}(0)$:

$$U_{1,k+1}(z) = \frac{A_1(z)}{z}\left(((1-\alpha)z + \alpha)U_{1,k}(z) + \alpha(z-1)U_{1,k}(0)\right) \; . \tag{1.29}$$

Our goal is to obtain the stationary distribution of $\{u_{1,k}\}_{k\in\mathbb{N}}$. Therefore, we define $p_1(i)$ and $U_1(z)$ as

$$p_1(i) \triangleq \lim_{k\to\infty} \Pr[u_{1,k} = i] \tag{1.30}$$

$$U_1(z) \triangleq \lim_{k\to+\infty} U_{1,k}(z) \tag{1.31}$$

$$= \sum_{i=0}^{\infty} p_1(i) z^i \; . \tag{1.32}$$

Taking the limit for $k \to +\infty$ in (1.29) and solving for $U_1(z)$ yields

$$U_1(z) = U_1(0)\frac{\alpha(z-1)A_1(z)}{z - A_1(z)(\alpha + (1-\alpha)z)} \; . \tag{1.33}$$

There is one quantity yet to be determined, namely the constant $U_1(0)$. This constant can be found from the normalization condition $U_1(1) = 1$, yielding

$$U_1(0) = 1 - \frac{\lambda_1}{\alpha} \; , \tag{1.34}$$

where we have used l'Hôpital's rule and (1.26). We finally get the following expression:

$$U_1(z) = \frac{(\alpha - \lambda_1)(z-1)A_1(z)}{z - A_1(z)(\alpha + (1-\alpha)z)} \; . \tag{1.35}$$

Theoretically, this PGF contains all information of the system content of type-1 customers. For example, we can compute relatively easy the mean and the variance of the number of type-1 customers. The mean is given by

$$\mathsf{E}[u_1] = \left.\frac{dU_1(z)}{dz}\right|_{z=1} \tag{1.36}$$

$$= \frac{2\lambda_1(1-\lambda_1) + A_1''(1)}{2(\alpha-\lambda_1)} \; . \tag{1.37}$$

The stability condition for the Bernoulli model (cf. [40, Ch. 3.2], Equation (3.71)) is given by

$$\lambda_1 < \alpha \; . \tag{1.38}$$

The stability condition is intuitively clear. The average number of customers entering queue-1 should be strictly less than the average number of customers that can be served from queue-1, per time-slot. Further notice that equation (1.34) shows that the stability condition is nothing else than the prerequisite that there is a positive probability that, at the beginning of a slot, there are no type-1 customers in the system. It can also easily be seen that the stability condition is incorporated in expression (1.37), since the mean number of type-1 customers increases to infinity when λ_1 approaches its critical value α.

Considering only type-2 customers, it is now easily seen that this is equivalent to a Bernoulli model with parameter $1-\alpha$ and arrival PGF $A_2(z)$. From the previous analysis we obtain, mutatis mutandis, the PGF $U_2(z)$ describing the stationary type-2 system content:

$$U_2(z) = \frac{(1-\alpha-\lambda_2)(z-1)A_2(z)}{z - A_2(z)(1-\alpha+\alpha z)} \; . \tag{1.39}$$

The mean type-2 system content is given by

$$\mathsf{E}[u_2] = \frac{2\lambda_2(1-\lambda_2) + A_2''(1)}{2(1-\alpha-\lambda_2)} \; . \tag{1.40}$$

The stability condition for the type-2 system content reads

$$\lambda_2 < 1-\alpha \; . \tag{1.41}$$

The analysis of a single queue with random server interruptions and single slot times is carried out in [40, Ch. 3.2] in a much more general setting. In this book it is assumed that the series of consecutive "available-periods" as well as the series of consecutive "interrupted-periods" share common general distributions. The only restriction is that the common probability generating function of the available periods must be a rational function. In the literature, queueing models with server interruptions are sometimes also referred to as systems with server

vacations or server breakdowns [45]. Queueing models with server interruptions can for instance be used to evaluate the performance of the individual queues in queueing systems with multiple queues sharing a common server such as priority systems or polling systems, see e.g. [46, 47] and [17, Sect. 5].

1.6 Goals and outline

Throughout this dissertation we study the *joint* distribution of the stationary system contents of a rudimentary, but conceptual, discrete-time two-class queueing system with a probabilistic scheduling discipline that does not depend on the system contents, which up till now has not been investigated in discrete-time queueing theory. Since we make extensive use of probability generating functions, the primary goal of this research is

Goal: To obtain an exact closed-form expression for the joint PGF of the stationary system contents, as a function of the input parameters.

From the joint PGF, every numerical characteristic of interest can be computed (in theory). However, in this dissertation we focus on finding expressions for the covariance of the two system contents and for the probability mass function of the total system content. The reason is that these quantities are both interesting and cannot be obtained from the single-queue analysis from Section 1.5.

Analysis of two-class queues is (obviously) more difficult than the analysis of single queues. In order to calculate most numerical characteristics, the joint distribution of the number of type-1 and type-2 customers in the queueing system has to be found. When the capacities are infinite, as it is the case in this dissertation, the domain of the joint distribution is unbounded in more than one dimension, turning an easy problem for the one-dimensional case into a problem that is almost infeasible to tackle exactly in the two-dimensional case. The analysis of this kind of models is considered to be notoriously hard [48]. The most well-developed analytical method for this kind of problems is the so-called boundary value method [49]. While this method provides an *exact* expression for the joint PGF, it is not a *closed-form* expression in the sense that it typically contains functions that most of the time have to be computed numerically through solving singular integral equations. Moreover, we emphasize that obtaining the probability mass function from this PGF will be a cumbersome task if the expression for the joint PGF is not in closed-form. Therefore, we downsize our primary goal to the following goal:

Goal I: To investigate which arrival processes give rise to an exact closed-form expression of the joint PGF and the joint pmf.

The study of this goal is carried out in Chapter 2 'Exact analysis: specific arrival distributions', in which we find three specific cases for the pmf $a(i,j)$ $(i,j \geq 0)$ describing the numbers of arrivals within a slot. The chapter is first introduced with the problem for general arrival processes and an extensive discussion of the analytical solution methods to this class of problems. Hereafter, three specific cases for $a(i,j)$ are discussed and analyzed. The emphasis is put on *how* the solution is obtained, rather than on the final result itself. However, we do provide explicit expressions for the most interesting numerical characteristics and discuss the influence of the scheduling discipline on some of these characteristics.

The urge for closed-form expressions can be questioned. Therefore we point out some of the main benefits of closed-form expressions. First and foremost, a closed-form expression for the solution of a problem involving many parameters is the most desirable one since the solution for any choice of parameters is then instantly obtained by just substituting the parameter values into the expression. Since closed-form expressions are easy to implement, they are the most suitable to investigate the sensitivity of input parameters on the output and to do optimization. Furthermore, closed-form expressions for physical quantities provide more insight compared to a complicate (whether or not exact) expression. Finally, closed-form expressions are also easier to present to a broader audience which may be primarily interested in the final results (and not how those results were obtained).

From the vast literature concerning similar two-class queueing models and based on our own experience, it seems very unlikely that a closed-form expression for the joint PGF exists for a general arrival PGF $A(z_1, z_2)$. Hence, the next step is to find closed-form expressions which are approximations for the problem at hand. There already exist well-developed approximation techniques for one-dimensional queueing problems. Hence, the second goal of this dissertation is:

Goal II: To study and apply well-known one-dimensional approximation/asymptotic techniques to the two-dimensional problem at hand.

The second goal of this dissertation is studied in Chapters 3-5. In Chapter 3, the goal is to obtain asymptotic formulas for the joint pmf of the stationary system contents. This will be accomplished using the theory of dominant singularities. Chapter 4 focuses on a method to approximate the joint PGF by a simple rational function, using the results from Chapter 3. Chapter 5 focuses on the joint heavy-traffic limit for the symmetric queueing model. Both singularity analysis (Chapter 3) and heavy-traffic analysis (Chapter 5) are well-known and among the most popular approximation techniques for one-dimensional queueing models. However, at present, the approximation technique we propose in Chapter 4 is actually not a state-of-the-art technique. While this novel technique is not accurate for every choice of parameters, it can pave the way

to new simpler and efficient approximation techniques.

In Chapter 6 we summarize the findings of this dissertation and address some interesting directions for future research.

Finally, we remark that the model studied in this dissertation can be easily simulated using a computer program. Confidence intervals for any numerical characteristic of interest can be obtained via this way. But notice that each specific choice of the input parameters, i.e. in our case α and $a(i,j)$ $(i,j \geq 0)$, requires a separate simulation run. While these simulation runs are nowadays very fast for most numerical characteristics, experience learns that simulation can be very time consuming if one is interested in the accurate estimation of the probability of rare events (such as joint tail probabilities). This is especially true if one wants to estimate rare events for numerous sets of parameters. Hence, for the case of estimating rare events, the analytical method certainly beats the simulation method.

1.7 Publications

The research conducted during the doctoral research has resulted in a number of publications. Most of the following publications have provided the material for this dissertation, aside from [50] and [51], that deal with a priority retrial queue with constant retrial policy. Although the model studied in [50] and [51] can also be interpreted as a two-class queueing model, we have decided to not include it in order to keep this dissertation as self-contained as possible.

1.7.1 Publications in international journals

1. A. Devos, J. Walraevens, T. Phung-Duc, H. Bruneel, Analysis of the queue lengths in a priority retrial queue with constant retrial policy, *Journal of Industrial & Management Optimization* 16(6), p. 2813-2842, 2020.

2. A. Devos, J. Walraevens, D. Fiems, H. Bruneel, Analysis of a discrete-time two-class randomly alternating service model with Bernoulli arrivals, *Queueing Systems* 96(1), p. 133-152, 2020.

3. A. Devos, J. Walraevens, D. Fiems and H. Bruneel, Heavy-Traffic Comparison of a Discrete-Time Generalized Processor Sharing Queue and a Pure Randomly Alternating Service Queue, *Mathematics* 9(21), Article 2723, no. of pages: 25, 2021.

4. A. Devos, J. Walraevens, D. Fiems and H. Bruneel, Approximations for the performance evaluation of a discrete-time two-class queue with an alternating service discipline, *Annals of Operations Research* 310(2), p. 477-503, 2022.

1.7.2 Papers in Proceedings of International Conferences

1. A. Devos, J. Walraevens, H. Bruneel, A priority retrial queue with constant retrial policy, *Proceedings of the 13th International Conference on Queueing Theory and Network Applications*, QTNA 2018 (Tsukuba, 25-27 July 2018), Lecture Notes in Computer Science, 2018, vol. 10932, pp. 3-21. Edit.: Y. Takahashi, T. Phung-Duc, S. Wittevrongel, W. Yue.

2. A. Devos, D. Fiems, J. Walraevens, H. Bruneel, An Approximate Analysis of a Bernoulli Alternating Service Model, *Proceedings of the 14th International Conference on Queueing Theory and Network Applications*, QTNA 2019 (Ghent, 27-29 August 2019), Lecture Notes in Computer Science, 2019, vol. 11688, pp. 314-329. Edit.: T. Phung-Duc, S. Kasahara, S. Wittevrongel.

1.7.3 Abstracts

1. A. Devos, J. Walraevens, D. Fiems, H. Bruneel, Heavy-traffic limit for a discrete-time two-class single server queueing model, *Abstracts of the 31st European Conference on Operational Research*, EURO 2021 (Athens, online, 11-14 July 2021), p. 249.

2

Exact analysis: specific arrival distributions

In this chapter, we derive a functional equation for the joint PGF of the stationary system contents in our queueing model as described in Section 1.4. This functional equation will play a key role throughout this dissertation. Various special cases from the perspective of the nature of the joint arrival distribution, are analyzed in detail in this chapter.

In our paper [52], we have obtained the joint PGF of the system contents in the special case that the numbers of type-1 and type-2 arrivals per time slot constitute two mutually independent sequences of i.i.d. random variables with common Bernoulli distribution. The analysis of Section 2.3 runs mainly parallel as in our contribution [52]. In addition, we also study two other arrival processes in this chapter. More specifically, we consider the case where the arrivals in both queues constitute two identical sequences of i.i.d. random variables with common Bernoulli distribution; and the case when the arrivals to the complete system constitute a sequence of i.i.d. random variables with common geometric distribution, which are randomly routed to one of the two queues.

The outline of the rest of this chapter is as follows. We give an overview of some literature on the exact analysis of two-dimensional queueing models via generating functions in Section 2.1. In particular, we focus on the -difficult to analyze- class of queueing models that can be solved using the theory of boundary-value problems for analytic functions. In Section 2.2 we establish the functional equation for the joint PGF of the stationary system contents. Section 2.3 is devoted to the special case of two independent Bernoulli arrival processes in the two queues. In section 2.4, we consider the case that the arrivals to the two queues are identical and Bernoulli distributed. Further, in section 2.5 we consider a single stream of geometrically distributed arrivals and probabilistic routing to the two queues. Finally, some conclusions are drawn in Section 2.6.

2.1 Exact analysis of two-dimensional queueing models

The analysis of queueing models involving two queues is substantially different from the analysis of a classical single-queue-single-server model. This is because these models often give rise to the problem of solving a functional equation of the form

$$\begin{aligned} K(z_1,z_2)\Phi(z_1,z_2) =& L_1(z_1,z_2)\Phi(z_1,0) + L_2(z_1,z_2)\Phi(0,z_2) \\ &+ L_3(z_1,z_2)\Phi(0,0)\ , \quad |z_1|\leq 1, |z_2|\leq 1; \end{aligned} \tag{2.1}$$

where K, L_j $(j = 1,2,3)$ are known functions and $\Phi(z_1,z_2)$ is the unknown function which represents the joint PGF of the numbers of customers in both queues, in steady state. Solving this equation requires finding the partial PGFs $\Phi(z_1,0)$ and $\Phi(0,z_2)$, which is the non-straightforward objective of the analysis. It is worth mentioning that substitution of $\{z_1 = z, z_2 = 0\}$ or $\{z_1 = 0, z_2 = z\}$ into (2.1) always leads to the tautology "$0 = 0$". The crucial part of the analysis is studying the function K, which is referred to as the *kernel* of the functional equation. This is because whenever a zero $(\hat{z}_1,\hat{z}_2)$ of K lies inside the region of convergence of the PGF $\Phi(z_1,z_2)$, this relates $\Phi(\hat{z}_1,0)$ with $\Phi(0,\hat{z}_2)$. More precisely, let Λ denote the set of zero-tuples of K which lie in the region of convergence of the joint PGF $\Phi(z_1,z_2)$. Then it is clear from (2.1) that for any $(\hat{z}_1,\hat{z}_2) \in \Lambda$,

$$L_1(\hat{z}_1,\hat{z}_2)\Phi(\hat{z}_1,0) + L_2(\hat{z}_1,\hat{z}_2)\Phi(0,\hat{z}_2) + L_3(\hat{z}_1,\hat{z}_2)\Phi(0,0) = 0\ ,$$

which gives us an equation in terms of $\Phi(\hat{z}_1,0)$ and $\Phi(0,\hat{z}_2)$.

In the pioneering paper [53], it is shown that $\Phi(z_1,0)$ as $\Phi(0,z_2)$ can be found as the solution of a boundary-value problem for analytic functions. In [53], two parallel M/M/1 queues with coupled service rates are analyzed. Because the approach as per [53] needs an explicit expression of the kernel K, the approach is limited to kernels of sufficiently simple type (which corresponds typically to Markovian assumptions). Cohen and Boxma [49] investigated if the *boundary-value approach* can be applied when the kernel has a more general character. For instance, for continuous-time queueing models, they considered the so-called Poisson kernel, i.e.,

$$K(z_1,z_2) = z_1z_2 - B^*(\lambda_1(1-z_1) + \lambda_2(1-z_2))\ ,$$

with $B^*(\cdot)$ the LST of a non-negative continuous random variable (typically the service times of the customers), and λ_j the rates of two independent Poissonian arrival streams. Cohen and Boxma showed that in case of a Poisson kernel, the problem of determining $\Phi(z_1,0)$ and $\Phi(0,z_2)$ from (2.1) can also be reduced to a boundary-value problem.

A myriad of queueing models have been analyzed by means of the boundary-value approach. Examples can be found in [54–56]. A detailed study of the

boundary-value approach is presented in the classic textbooks [49, 57]. While the theory behind the boundary-value approach was developed during the late 70s and early 80s, it is still frequently applied to solve functional equations like (2.1). Examples of discussions and applications of the theory of boundary-value problems to recent queueing models can be found in [38, 58–68].

Solving (2.1) is considered to be a notoriously hard problem, because the standard state-of-the-art methodology exists of treating it as a boundary-value problem. Moreover, in practice, cumbersome numerical work is necessary afterwards to obtain actual numerical characteristics of the queueing model, such as the probability that there are more than a specified number of customers in the (total) system, the correlation between the numbers of customers in both queues, etc. This is because in order to evaluate the expressions obtained through the boundary-value method, a numerical solution of singular integral equations is required. See for example the papers [58, 69] where this numerical approach is carried out.

As mentioned in the introductory chapter, we are interested in closed-form expressions. The boundary-value approach (although it is powerful) seems to be limited in that perspective. Up till now, only a few queueing models can be analyzed exactly via generating functions, avoiding the boundary-value technique. We discuss three two-dimensional queueing models which have been well-studied in this context.

(A) The continuous-time shortest queue: consists of two queues with two dedicated homogeneous servers. There is a single Poisson arrival process. A newly arrived customer joins the shortest of the two queues. If the queue lengths are equal, the customer joins either with probability $\frac{1}{2}$.

(B) The discrete-time 2×2 switch queue: consists of two queues with two dedicated homogeneous servers. The arrival process is characterized by means of a sequence of i.i.d. discrete random vectors with common PGF:

$$A(z_1, z_2) = \left(1 - \lambda + \tfrac{\lambda}{2}(z_1 + z_2)\right)^2 .$$

(C) The continuous-time fork-join queue: consists of two queues with two dedicated heterogeneous servers. Customers arrive in batches of size two according to a Poisson arrival process. Two newly arrived customers from the same batch each enter a different queue. The two customers that arrived in the same batch leave the system only if both customers have been served.

These three models do not exhibit a product-form solution[1], because the arrivals at the two queues are correlated. Among the classic papers analyzing model (A) are [71] and [72]. Using generating functions, it is proven that

[1] Product-form basically means that the system contents are independent and one can just multiply the marginal distributions [70].

the partial PGFs $\Phi(z_1, 0)$ and $\Phi(0, z_2)$ are meromorphic functions. In [72], this is accomplished by finding a suitable parametrization of the manifold $\{(z_1, z_2) : K(z_1, z_2) = 0\}$. The approach as per [71] is comparable to the one in [72], but the latter is more elaborated in detail. Because the essential element in these two approaches is a parametrization, this method goes by the name of *the uniformization method*. This method can also be used for models (B) and (C). For example, in model (B) it is also shown that the partial PGFs are meromorphic by means of the uniformization method [73]. On the other hand, in [74] and [75] it is shown for model (A) and model (B), respectively, that meromorphicity can also be established from the functional equation directly, i.e., without the need of a parametrization of $\{(z_1, z_2) : K(z_1, z_2) = 0\}$. The joint pmf of the system contents in (A) and (B) can be expressed as infinite sums, which in practice require truncation to compute. The analysis of the asymmetric versions of (A) and (B) have been accomplished in [76, 77]. In these papers, the analysis does not make use of a parametrization, but directly uses analytic continuation as per [75]. Further, the same author shows in [78] that his method is also applicable to a larger class of queueing models. Namely, to the class of nearest-neigbour random walks in $\{0, 1, \ldots\}^2$ with no one-step displacements to the North, North-East and the East.

Model (C) as described above was introduced in [79] and the model was later generalized in [80] by adding an extra Poisson arrival stream without batches. The uniformization method is used in [79] to obtain the unknown partial PGFs. While in models (A) and (B) the parametrization is accomplished by a pair of rational functions, model (C) requires a pair of elliptic functions, say $z_1 = v(t)$ and $z_2 = w(t)$. In [79], it is shown that $\Phi(v(t), 0)$ and $\Phi(0, w(t))$ admit analytic continuation as meromorphic functions. We emphasize that the analysis of (C) involves some ingenious applications of complex analysis.

The approach as per Section 2.3 is comparable to the one in [75] and [76]. More precisely, we will solve the functional equation using the notion of analytic continuation as in [75, 76]. However, since the kernel in our case is different from the ones in [75, 76], new difficulties arise. We emphasize that, for the case as in Section 2.3, we are unable to solve the the functional equation using the boundary-value approach, since we are unable to find the conformal mapping in closed form. The kernels that we encounter in Section 2.4 and Section 2.5 have the same structure as the one in [81]. Therefore the analysis of the kernel K in these sections is comparable to the one in [81]. The difference between [81] and Sections 2.4 and 2.5 is that the RHS of the functional equation does not have the same structure. Hence, at some point in the analysis, a different approach is used in Sections 2.4 and 2.5 (in comparison with [81]).

Finally, we want to draw attention to the compensation method [106]. This method provides an exact expression for the joint stationary distribution of certain two-dimensional random walks as an infinite sum of product-form terms. However, we note that this method is not based on PGFs. Some further details and references of the compensation method are discussed in Section 4.1.

2.2 A functional equation for $U(z_1, z_2)$

The purpose of this section is to show that the joint PGF of the stationary system contents of our model satisfies a functional equation like (2.1).

As introduced in Section 1.4, the number of type-j arrivals in slot k is denoted by $a_{j,k}$. The joint PGF of $a_{1,k}$ and $a_{2,k}$ is denoted by $A(z_1, z_2)$. Further, we defined the sequence $\{r_k\}$, which is a sequence of i.i.d. random variables with common Bernoulli distribution with parameter α.

Let $u_{j,k}$ $(j = 1, 2)$ denote the system content of type j at the beginning of slot k. The corresponding joint PGF of $u_{1,k}$ and $u_{2,k}$ is denoted by $U_k(z_1, z_2)$, i.e.

$$U_k(z_1, z_2) \triangleq \mathsf{E}[z_1^{u_{1,k}} z_2^{u_{2,k}}] \; . \tag{2.2}$$

The evolution of the system content from slot k to slot $k + 1$ is described by the following system equations:

$$u_{1,k+1} = (u_{1,k} - r_k)^+ + a_{1,k} \; , \tag{2.3}$$

$$u_{2,k+1} = (u_{2,k} - 1 + r_k)^+ + a_{2,k} \; . \tag{2.4}$$

The system equations follow from the fact that if $r_k = 1$, type-1 customers can be served. In this case, type-2 customer cannot be served, even if there were no type-1 customers at the beginning of slot k. Analogously, we have the symmetric case when $r_k = 0$.

From the system equations we obtain a relation for the joint PGF $U_{k+1}(z_1, z_2)$ of the number of type-1 and type-2 customers at the beginning of slot $k + 1$ and the joint PGF $U_k(z_1, z_2)$ of the number of type-1 and type-2 customers at the beginning of slot k:

$$\begin{aligned}
U_{k+1}(z_1, z_2) &\triangleq \mathsf{E}[z_1^{u_{1,k+1}} z_2^{u_{2,k+1}}] \\
&= \mathsf{E}[z_1^{(u_{1,k}-r_k)^+ + a_{1,k}} z_2^{(u_{2,k}-1+r_k)^+ + a_{2,k}}] \\
&= (1-\alpha)\mathsf{E}[z_1^{u_{1,k}+a_{1,k}} z_2^{(u_{2,k}-1)^+ + a_{2,k}}] \\
&\quad + \alpha \mathsf{E}[z_1^{(u_{1,k}-1)^+ + a_{1,k}} z_2^{u_{2,k}+a_{2,k}}] \\
&= A(z_1, z_2) \left\{ (1-\alpha)\mathsf{E}[z_1^{u_{1,k}} z_2^{(u_{2,k}-1)^+}] + \alpha \mathsf{E}[z_1^{(u_{1,k}-1)^+} z_2^{u_{2,k}}] \right\} \\
&= A(z_1, z_2) \left\{ \frac{(1-\alpha)}{z_2} (U_k(z_1, z_2) + (z_2 - 1)U_k(z_1, 0)) \right. \\
&\quad \left. + \frac{\alpha}{z_1} (U_k(z_1, z_2) + (z_1 - 1)U_k(0, z_2)) \right\} \\
&= \frac{A(z_1, z_2)}{z_1 z_2} \{ [(1-\alpha)z_1 + \alpha z_2] U_k(z_1, z_2) \\
&\quad + (1-\alpha) z_1 (z_2 - 1) U_k(z_1, 0) + \alpha z_2 (z_1 - 1) U_k(0, z_2) \} \; .
\end{aligned} \tag{2.5}$$

Notice that

$$U_k(z_1,0) = \mathsf{E}[z_1^{u_{1,k}}\mathbf{1}\{u_{2,k}=0\}] \tag{2.6}$$

$$= \sum_{n=0}^{\infty} \Pr[u_{1,k}=n, u_{2,k}=0] z_1^n \ , \tag{2.7}$$

and

$$U_k(0,z_2) = \mathsf{E}[z_2^{u_{2,k}}\mathbf{1}\{u_{1,k}=0\}] \tag{2.8}$$

$$= \sum_{n=0}^{\infty} \Pr[u_{1,k}=0, u_{2,k}=n] z_2^n \ , \tag{2.9}$$

by definition. Since we are interested in the joint stationary distribution of $u_{1,k}$ and $u_{2,k}$, we define

$$p(i,j) \triangleq \lim_{k\to\infty} \Pr[u_{1,k}=i, u_{2,k}=j] \ , \tag{2.10}$$

$$U(z_1,z_2) \triangleq \lim_{k\to\infty} U_k(z_1,z_2)$$

$$= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p(i,j) z_1^i z_2^j \ . \tag{2.11}$$

Assuming that the system reaches a steady state, then both functions $U_k(z_1,z_2)$ and $U_{k+1}(z_1,z_2)$ converge to the common limit function $U(z_1,z_2)$. As a result, by taking the limit $k\to\infty$ in equation (2.5) we obtain the following functional equation for $U(z_1,z_2)$:

$$K(z_1,z_2)U(z_1,z_2) = A(z_1,z_2)$$
$$\times [(1-\alpha)(z_2-1)z_1 U(z_1,0) + \alpha(z_1-1)z_2 U(0,z_2)] \ , \tag{2.12}$$

where we defined

$$K(z_1,z_2) \triangleq z_1 z_2 - [(1-\alpha)z_1 + \alpha z_2] A(z_1,z_2) \ . \tag{2.13}$$

Throughout the rest of this dissertation, we will use the following notation

$$K^{(j)}(x,y) \triangleq \left.\frac{\partial K(z_1,z_2)}{\partial z_j}\right|_{z_1=x,z_2=y} \ , \tag{2.14}$$

$$K^{(ij)}(x,y) \triangleq \left.\frac{\partial^2 K(z_1,z_2)}{\partial z_i \partial z_j}\right|_{z_1=x,z_2=y} \ , \quad i,j=1,2. \tag{2.15}$$

Finally, the joint PGF can be expressed as

$$U(z_1,z_2) = \frac{A(z_1,z_2)[(1-\alpha)(z_2-1)z_1 U(z_1,0) + \alpha(z_1-1)z_2 U(0,z_2)]}{z_1 z_2 - [(1-\alpha)z_1 + \alpha z_2]A(z_1,z_2)} \ . \tag{2.16}$$

There are two unknown functions yet to be determined in the right-hand side of (2.16), namely the functions $U(z_1,0)$ and $U(0,z_2)$.

2.3 Independent Bernoulli arrivals in the two queues

In this section, we assume that the random variables $a_{1,k}$ and $a_{2,k}$ constitute two independent sequences of independent and identically Bernoulli distributed random variables. As a consequence, the joint PGF $A(z_1, z_2)$ is given by

$$A(z_1, z_2) = (1 - \lambda_1 + \lambda_1 z_1)(1 - \lambda_2 + \lambda_2 z_2) \; . \tag{2.17}$$

This is the simplest model in the case that the numbers of arrivals of type-1 and type-2 customers are independent, i.e. $A(z_1, z_2) = A_1(z_1)A_2(z_2)$. The results of this section are based on our contribution [52]. As we will show, the two unknown partial generating functions $U(z, 0)$ and $U(0, z)$ admit analytic continuation as rational functions with two poles each.

2.3.1 The marginal distributions $p_1(n)$ and $p_2(n)$

In Chapter 1, Section 1.5, we showed that the marginal PGFs of the system contents can be easily obtained. From (2.17), it easily follows that

$$\begin{aligned} A_1(z) &= A(z, 1) \\ &= 1 - \lambda_1 + \lambda_1 z \; . \end{aligned}$$

Substituting this arrival PGF into (1.35) yields

$$U_1(z) = \frac{(\alpha - \lambda_1)(1 - \lambda_1 + \lambda_1 z)}{\alpha(1 - \lambda_1) - \lambda_1(1 - \alpha)z} \; , \tag{2.18}$$

in which we canceled the common factor $(z-1)$ in numerator and denominator. Notice that this PGF can also be deduced from (2.16) by substituting $\{z_1 = z, z_2 = 1\}$, i.e.

$$U_1(z) = U(z, 1) \; .$$

Recall that (cf. Definition (1.31)) $U_1(z)$ is defined by

$$U_1(z) = \sum_{i=0}^{\infty} p_1(i) z^i \; .$$

Hence, by writing down the Taylor series expansion of (2.18) and by equating coefficients, it follows that

$$\begin{aligned} p_1(0) &= 1 - \frac{\lambda_1}{\alpha} \; , \\ p_1(n) &= \frac{\alpha - \lambda_1}{\alpha(1 - \alpha)} \frac{1}{\tau_1^n}, \quad n \geq 1 \; , \end{aligned}$$

where τ_1 is the unique zero of the denominator in (2.18), given by

$$\tau_1 = \frac{\alpha}{1-\alpha}\frac{1-\lambda_1}{\lambda_1} \ . \tag{2.19}$$

Obviously, τ_1 satisfies the equation $K(z,1) = 0$.

We can obtain the PGF $U_2(z)$ describing the system content of type-2 customers in a similar way. It follows that

$$U_2(z) = \frac{(1-\alpha-\lambda_2)(1-\lambda_2+\lambda_2 z)}{(1-\alpha)(1-\lambda_2)-\alpha\lambda_2 z} \ . \tag{2.20}$$

Further, because

$$U_2(z) = \sum_{i=0}^{\infty} p_2(i) z^i \ ,$$

we obtain that

$$\begin{aligned} p_2(0) &= 1 - \frac{\lambda_2}{1-\alpha} \ , \\ p_2(n) &= \frac{1-\alpha-\lambda_2}{\alpha(1-\alpha)}\frac{1}{\tau_2^n}, \quad n \geq 1 \ , \end{aligned}$$

where τ_2 is the unique zero of the denominator in (2.20), given by

$$\tau_2 = \frac{1-\alpha}{\alpha}\frac{1-\lambda_2}{\lambda_2} \ . \tag{2.21}$$

Obviously, τ_2 satisfies the equation $K(1,z) = 0$.

Finally, we consider the marginal PGF $U_T(z)$ of the total number of customers in both queues together. A correct expression is obtained by choosing $\{z_1 = z, z_2 = z\}$ in (2.16):

$$\begin{aligned} U_T(z) &\triangleq U(z,z) \\ &= \frac{(1-\lambda_1+\lambda_1 z)(1-\lambda_2+\lambda_2 z)}{(1-\lambda_1)(1-\lambda_2)-\lambda_1\lambda_2 z}[(1-\alpha)U(z,0)+\alpha U(0,z)] \ , \end{aligned} \tag{2.22}$$

which, unfortunately, contains the unknown terms $U(z,0)$ and $U(0,z)$ again. We briefly take a look at the dominant singularities of (2.22). The dominant singularity of $U_T(z)$ is either the dominant singularity of $U(z,0)$, the dominant singularity of $U(0,z)$ or the zero of the denominator τ_T, given by

$$\tau_T = \frac{(1-\lambda_1)(1-\lambda_2)}{\lambda_1\lambda_2} \ . \tag{2.23}$$

Obviously, τ_T satisfies the equation $K(z,z) = 0$. Further, it is worth noting that $\tau_T = \tau_1\tau_2$.

2.3.2 Areas of convergence

For definiteness we recall definitions (2.7) and (2.10). The boundary function $U(z,0)$ is defined by

$$U(z,0) = \sum_{i=0}^{\infty} p(i,0) z^i \ , \tag{2.24}$$

the power series of the horizontal boundary probabilities. Similarly, the boundary function $U(0,z)$ is defined by

$$U(0,z) = \sum_{i=0}^{\infty} p(0,i) z^i \ , \tag{2.25}$$

the power series of the vertical boundary probabilities. We now investigate for which values of z these two infinite series converge. To accomplish this, we observe that for every $i \in \mathbb{N}$

$$\begin{aligned} p(i,0) &\leq p(i,0) + p(i,1) + p(i,2) + \ldots \\ &= \sum_{j=0}^{\infty} p(i,j) \\ &= \sum_{j=0}^{\infty} \lim_{k\to\infty} \Pr[u_{1,k} = i, u_{2,k} = j] \\ &= \lim_{k\to\infty} \Pr[u_{1,k} = i] \\ &= p_1(i) \ . \end{aligned}$$

Hence, the radius of convergence of $U_1(z)$ is a lower bound for the radius of convergence of $U(z,0)$. Analogously, the radius of convergence of $U_2(z)$ is a lower bound for the radius of convergence of $U(0,z)$. From Section 2.3.1, it easily follows that the radius of convergence of $U_1(z)$ and $U_2(z)$ is given by τ_1 and τ_2, respectively. Consequently, $U(z,0)$ and $U(0,z)$ have to be analytic in $|z| < \tau_1$ and $|z| < \tau_2$, respectively.

Next, we investigate the joint PGF $U(z_1,z_2)$. For any z_2 with modulus smaller than or equal to 1, we have

$$\begin{aligned} |U(z_1,z_2)| &\leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p(i,j)|z_1|^i|z_2|^j \\ &\leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p(i,j)|z_1|^i \\ &= \sum_{i=0}^{\infty} p_1(i)|z_1|^i \\ &= U_1(|z_1|) \ . \end{aligned}$$

From the fact that the radius of convergence of $U_1(z)$ is given by τ_1, we further obtain that

$$|U(z_1, z_2)| < \infty, \quad \text{if } |z_1| < \tau_1, |z_2| \leq 1 \ .$$

For reasons of symmetry, we can similarly prove that

$$|U(z_1, z_2)| < \infty, \quad \text{if } |z_1| \leq 1, |z_2| < \tau_2 \ .$$

Let us define two regions in $\mathbb{C}^2$:

$$\Omega_1 = \{(z_1, z_2) \mid |z_1| < \tau_1, |z_2| \leq 1\} \ , \tag{2.26}$$
$$\Omega_2 = \{(z_1, z_2) \mid |z_1| \leq 1, |z_2| < \tau_2\} \ . \tag{2.27}$$

We have thus shown that for $(z_1, z_2) \in \Omega_1 \cup \Omega_2$, the double power series expansion of $U(z_1, z_2)$ (cf. (2.11)) converges and it is consequently finite in this region.

2.3.3 Analysis of the kernel $K(z_1, z_2)$

As mentioned in the general introduction of this chapter, a central role in the analysis of the functional equation (2.1) is played by the kernel

$$\begin{aligned} K(z_1, z_2) &= z_1 z_2 - ((1-\alpha)z_1 + \alpha z_2)A(z_1, z_2) \\ &= z_1 z_2 - ((1-\alpha)z_1 + \alpha z_2)(1 - \lambda_1 + \lambda_1 z_1)(1 - \lambda_2 + \lambda_2 z_2) \ . \end{aligned} \tag{2.28}$$

Since the kernel is at most quadratic in z_1 and z_2, we can make a detailed study of the zeros of the kernel K.

The function $Y_1(z)$

We can rewrite $K(z_1, z_2)$ as follows

$$\begin{aligned} K(z_1, z_2) = &- \alpha\lambda_2 A_1(z_1) z_2^2 + (z_1 - A_1(z_1)[\alpha(1-\lambda_2) + (1-\alpha)\lambda_2 z_1])z_2 \\ &- (1-\alpha)(1-\lambda_2) z_1 A_1(z_1) \ , \end{aligned} \tag{2.29}$$

with $A_1(z) = A(z, 1) = 1 - \lambda_1 + \lambda_1 z$. Observe that $K(z_1, z_2)$ is for each $z_1 \neq -\frac{1-\lambda_1}{\lambda_1}$ a polynomial of degree 2. The corresponding discriminant, denoted by $D_Y(z_1)$, is given by

$$D_Y(z_1) = \{z_1 - A_1(z_1)[\alpha(1-\lambda_2) + (1-\alpha)\lambda_2 z_1]\}^2 - 4\alpha(1-\alpha)\lambda_2(1-\lambda_2) z_1 A_1^2(z_1) \ .$$

It is easily verified that $D_Y(z)$ is a polynomial of degree 4. Let us denote the zeros of the polynomial $D_Y(z)$ by x_1, x_2, x_3 and x_4.

Lemma 2.1. *The zeros of $D_Y(z)$ are real, moreover*

$$0 < x_1 < x_2 < 1 < \tau_T < x_3 < x_4 < \infty \ .$$

Proof. We define

$$h(x) \triangleq x - A_1(x)[\alpha(1-\lambda_2) + (1-\alpha)\lambda_2 x] \,,$$

which is a polynomial of degree 2. It follows that $D_Y(x) = h^2(x) - 4\alpha(1-\alpha)\lambda_2(1-\lambda_2)xA_1^2(x)$. Because

$$\begin{aligned} h(0) &= -A_1(0)\alpha(1-\lambda_2) \,, \\ h(1) &= 1 - \alpha - \lambda_2 + 2\alpha\lambda_2 \,, \end{aligned}$$

one can easily verify that $h(0) < 0$ and $h(1) > 0$ (using the stability condition $\lambda_2 < 1-\alpha$). Furthermore, $\lim_{x\to\infty} h(x) < 0$. Hence, $h(x)$ must have a zero in $]0,1[$, and in $]1,\infty[$. Moreover we have that the latter zero is larger than τ_T because

$$\begin{aligned} h(\tau_T) &= \tau_T(\alpha - 2\alpha\lambda_1 + \lambda_1) \\ &= \tau_T(\alpha(1-\lambda_1) + \lambda_1(1-\alpha)) \\ &> 0 \,. \end{aligned}$$

It now easily follows that D_Y is strictly negative in the two zeros of $h(x)$. If we combine these observations with $D_Y(0) > 0$, $D_Y(1) > 0$, $D_Y(\tau_T) > 0$ and $\lim_{x\to\infty} D_Y(x) > 0$, we can conclude that $D_Y(x)$ changes sign twice in $[0,1]$ and twice in $[\tau_T, +\infty[$. Hence, the zeros of $D_Y(z)$ are real and we have necessarily that $0 < x_1 < x_2 < 1 < \tau_T < x_3 < x_4 < \infty$. □

The function $D_Y(z)$ can thus be factorized as

$$D_Y(z) = (1-\alpha)^2\lambda_1^2\lambda_2^2(z-x_1)(z-x_2)(z-x_3)(z-x_4) \,.$$

Further we define,

$$\Delta(z) \triangleq (1-\alpha)\lambda_1\lambda_2\sqrt{z-x_1}\sqrt{z-x_2}\sqrt{z-x_3}\sqrt{z-x_4} \,, \tag{2.30}$$

with $\sqrt{\cdot}$ the principal value of the square root. Next we define the following two functions

$$Y_1(z) \triangleq \frac{z - A_1(z)[\alpha(1-\lambda_2) + (1-\alpha)\lambda_2 z] + \Delta(z)}{2\alpha\lambda_2 A_1(z)} \tag{2.31}$$

$$Y_2(z) \triangleq \frac{z - A_1(z)[\alpha(1-\lambda_2) + (1-\alpha)\lambda_2 z] - \Delta(z)}{2\alpha\lambda_2 A_1(z)} \tag{2.32}$$

One easily sees that $K(z, Y_1(z)) = 0$, $K(z, Y_2(z)) = 0$ $\forall z \in \mathbb{C} \setminus \{-\frac{1-\lambda_1}{\lambda_1}\}$, and

$$Y_1(z)Y_2(z) = \tau_2 z, \quad \forall z \in \mathbb{C} \setminus \{-\tfrac{1-\lambda_1}{\lambda_1}\} \,. \tag{2.33}$$

The denominators in (2.31) and (2.32) have one unique zero, namely $z = -\frac{1-\lambda_1}{\lambda_1}$. The numerator in (2.31) vanishes for $z = -\frac{1-\lambda_1}{\lambda_1}$. Therefore Y_1 is

analytic in $\mathbb{C} \setminus \{[x_1, x_2] \cup [x_3, x_4]\}$. Y_1 has one zero, namely $z = 0$, while Y_2 has no zeros. Further, we have that

$$\frac{Y_1(z)}{Y_2(z)} \to 0, \quad \text{as } z \to \infty ,$$

and $|Y_1(z)| = |Y_2(z)|$ for $z \in [x_1, x_2] \cup [x_3, x_4]$ (since $Y_1(z)$ and $Y_2(z)$ are complex conjugate in this interval). Applying the maximum principle [9, Th. 12, page 134] to the function $\frac{Y_1(z)}{Y_2(z)}$ yields

$$|Y_1(z)| < |Y_2(z)|, \quad z \in \mathbb{C} \setminus \{[x_1, x_2] \cup [x_3, x_4]\} . \tag{2.34}$$

Finally, based on Rouché's theorem, we deduce two bounds on the modulus of $Y_1(z)$ that will turn out useful in the analysis of the functional equation.

Lemma 2.2.

1. *If* $1 < |z| < \tau_1$, *we have that* $|Y_1(z)| < 1$
2. *If* $1 < |z| < \tau_T$, *then* $|Y_1(z)| < \tau_2$.

Proof. (1): First notice that

$$x - ((1-\alpha)x + \alpha)(1 - \lambda_1 + \lambda_1 x) = K(x, 1) > 0 \quad \text{for} \quad x \in]1, \tau_1[, \tag{2.35}$$

since 1 and τ_1 are the two zeros of the quadratic function $x \mapsto K(x, 1)$ and $\lim_{x \to +\infty} K(x, 1) = -\infty$.

Let the complex value z_1 be fixed, $1 < |z_1| < \tau_1$. On $|z_2| = 1$, we have

$$\begin{aligned}
&|((1-\alpha)z_1 + \alpha z_2)A(z_1, z_2)| \\
&= |((1-\alpha)z_1 + \alpha z_2)(1 - \lambda_1 + \lambda_1 z_1)(1 - \lambda_2 + \lambda_2 z_2)| \\
&\leq ((1-\alpha)|z_1| + \alpha|z_2|)(1 - \lambda_1 + \lambda_1|z_1|)(1 - \lambda_2 + \lambda_2|z_2|) \\
&= ((1-\alpha)|z_1| + \alpha)(1 - \lambda_1 + \lambda_1|z_1|) .
\end{aligned}$$

On the other hand, we have that $|z_1 z_2| = |z_1|$ on the circle $|z_2| = 1$. Because of (2.35), we have the inequality

$$\begin{aligned}
|((1-\alpha)z_1 + \alpha z_2)A(z_1, z_2)| &\leq ((1-\alpha)|z_1| + \alpha)(1 - \lambda_1 + \lambda_1|z_1|) \\
&< |z_1| .
\end{aligned}$$

By virtue of Rouché's theorem, the number of zeros of $z_1 z_2$ inside $|z_2| \leq 1$ is then the same as the number of zeros of $K(z_1, z_2)$. The former number is 1 (due to the trivial zero $z_2 = 0$). Hence, we have found that for fixed z_1, $1 < |z_1| < \tau_1$, the function $z_2 \mapsto K(z_1, z_2)$ has exactly one zero inside the unit disk, say $R(z_1)$. Since the only two zeros of $z_2 \mapsto K(z_1, z_2)$ are $Y_1(z_1)$ and $Y_2(z_1)$, $|Y_1(z_1)| < |Y_2(z_1)|$, we necessarily have that $R(z_1) = Y_1(z_1)$.

(2): Notice now that

$$x\tau_2-((1-\alpha)x+\alpha\tau_2)(1-\lambda_1+\lambda_1 x)(1-\lambda_2+\lambda_2\tau_2) = K(x,\tau_2) > 0 \text{ for } x \in]1,\tau_T[\ , \tag{2.36}$$

since 1 and τ_T are the two zeros of the quadratic function $x \mapsto K(x,\tau_2)$ and $\lim_{x\to+\infty} K(x,\tau_2) = -\infty$.

Let the complex value z_1 be fixed such that $1 < |z_1| < \tau_T$. On $|z_2| = \tau_2$, we have

$$\begin{aligned}
&|((1-\alpha)z_1+\alpha z_2)A(z_1,z_2)| \\
&= |((1-\alpha)z_1+\alpha z_2)(1-\lambda_1+\lambda_1 z_1)(1-\lambda_2+\lambda_2 z_2)| \\
&\le ((1-\alpha)|z_1|+\alpha|z_2|)(1-\lambda_1+\lambda_1|z_1|)(1-\lambda_2+\lambda_2|z_2|) \\
&= ((1-\alpha)|z_1|+\alpha\tau_2)(1-\lambda_1+\lambda_1|z_1|)(1-\lambda_2+\lambda_2\tau_2)\ .
\end{aligned}$$

On the other hand, we have that $|z_1 z_2| = |z_1|\tau_2$ on the circle $|z_2| = \tau_2$. The remainder of the proof is analogous to the proof of part (1), making use of Rouché 's theorem on the contour $|z_2| = \tau_2$ and (2.36). □

The function $X_1(z)$

We now consider $K(z_1,z_2)$ as a polynomial in the variable z_1 with $z_2 \neq -\frac{1-\lambda_2}{\lambda_2}$. The discriminant is again a polynomial of degree 4, given by

$$D_X(z_2) = \{z_2-A_2(z_2)[(1-\alpha)(1-\lambda_1)+\alpha\lambda_1 z_2]\}^2-4\alpha(1-\alpha)\lambda_1(1-\lambda_1)z_2A_2(z_2)^2\ .$$

The zeros of $D_X(z)$ are denoted by y_1, y_2, y_3 and y_4.

Lemma 2.3. *The zeros of $D_X(z)$ are real, moreover*

$$0 < y_1 < y_2 < 1 < \tau_T < y_3 < y_4 < \infty\ .$$

Proof. The proof is the same as the proof of Lemma 2.1 and is therefore omitted. □

Let us define

$$\Psi(z) \triangleq \alpha\lambda_1\lambda_2\sqrt{z-y_1}\sqrt{z-y_2}\sqrt{z-y_3}\sqrt{z-y_4}\ , \tag{2.37}$$

and two functions

$$X_1(z) \triangleq \frac{z-A_2(z)[(1-\alpha)(1-\lambda_1)+\alpha\lambda_1 z]+\Psi(z)}{2(1-\alpha)\lambda_1 A_2(z)} \tag{2.38}$$

$$X_2(z) \triangleq \frac{z-A_2(z)[(1-\alpha)(1-\lambda_1)+\alpha\lambda_1 z]-\Psi(z)}{2(1-\alpha)\lambda_1 A_2(z)} \tag{2.39}$$

It can be seen that $K(X_1(z),z) = 0$, $K(X_2(z),z) = 0\ \forall z \in \mathbb{C}\setminus\{-\frac{1-\lambda_2}{\lambda_2}\}$ and

$$X_1(z)X_2(z) = \tau_1 z, \quad \forall z \in \mathbb{C}\setminus\{-\tfrac{1-\lambda_2}{\lambda_2}\}\ . \tag{2.40}$$

One can then again verify that $X_1(z)$ is an analytic function for $z \in \mathbb{C}\backslash\{[y_1, y_2] \cup [y_3, y_4]\}$. Finally, using the maximum principle, one can show that

$$|X_1(z)| < |X_2(z)|, \quad z \in \mathbb{C} \setminus \{[y_1, y_2] \cup [y_3, y_4]\} . \tag{2.41}$$

Finally, we have the equivalent of Lemma 2.2 for the modulus of $X_1(z)$.

Lemma 2.4.

1. *If* $1 < |z| < \tau_2$, *we have that* $|X_1(z)| < 1$
2. *If* $1 < |z| < \tau_T$, *then* $|X_1(z)| < \tau_1$.

Proof. The proof is the same as the proof of Lemma 2.2 and is therefore omitted. □

2.3.4 Analytic continuation of $U(z, 0)$ and $U(0, z)$

We analytically continue the functions $U(z, 0)$ and $U(0, z)$. Such a continuation is unique and the extended functions restricted to $|z| < \tau_1$, $|z| < \tau_2$, coincide with the power series expansions (2.24) and (2.25). We will denote the analytic continuation also by $U(z, 0)$ and $U(0, z)$. The final result of this procedure will be that the (analytically continued) functions are rational functions with only two simple poles.

Continuation to $|z| < \tau_T$

We are now ready to continue analytically $U(z, 0)$ and $U(0, z)$ outside $|z| < \tau_1$ and $|z| < \tau_2$ respectively.

Theorem 2.1.

1. *The function* $U(z, 0)$ *can be analytically continued to* $\tau_1 \leq |z| < \tau_T$, $z \neq \tau_1$.
2. *The function* $U(0, z)$ *can be analytically continued to* $\tau_2 \leq |z| < \tau_T$, $z \neq \tau_2$.

Proof. (1): Because of Lemma 2.2 we have that $|Y_1(z)| < 1$ if $1 < |z| < \tau_1$, hence $U(z, Y_1(z))$ remains finite. Therefore, substituting $\{z_1 = z, z_2 = Y_1(z)\}$ into the functional equation (2.12) yields

$$(1 - \alpha)(Y_1(z) - 1)zU(z, 0) + \alpha(z - 1)Y_1(z)U(0, Y_1(z)) = 0, \quad 1 < |z| < \tau_1 ,$$

or

$$(1-\alpha)(Y_1(z)-1)zU(z, 0) = -\alpha(z-1)Y_1(z)U(0, Y_1(z)), \quad 1 < |z| < \tau_1 . \tag{2.42}$$

Both sides of this equation are analytic functions in $1 < |z| < \tau_1$ because $U(z,0)$, $Y_1(z)$ and $U(0,Y_1(z))$ are. According to Lemma 2.2, we have that $|Y_1(z)| < \tau_2$ whenever $1 < |z| < \tau_T$. Because of this bound, it follows that the RHS in (2.42) is an analytic function in the larger region $1 < |z| < \tau_T$. Hence, we can analytically continue $(1-\alpha)(Y_1(z)-1)zU(z,0)$ into the region $\tau_1 \leq |z| < \tau_T$ via (2.42). Because $(1-\alpha)(Y_1(z)-1)zU(z,0)$ is now analytic in $\tau_1 \leq |z| < \tau_T$, it follows that $U(z,0)$ is meromorphic in $\tau_1 \leq |z| < \tau_T$. The poles of $U(z,0)$ in $\tau_1 \leq |z| < \tau_T$ (if any) are the roots of $Y_1(z) - 1 = 0$.

Without using the expression for $Y_1(z)$, it can be shown that $z = 1$ and $z = \tau_1$ are the only zeros of $Y_1(z) - 1$. The proof goes as follows. First, the two zeros of $K(1,z)$ are $z = 1$ and $z = \tau_2$. By definition of Y_1 and Y_2, it holds that $Y_1(1) = 1$ and $Y_2(1) = \tau_2$. Secondly, the two zeros of $K(\tau_1,z)$ are $z = 1$ and $z = \tau_T$. Hence, it holds that $Y_1(\tau_1) = 1$ and $Y_2(\tau_1) = \tau_T$. Up till now, we have thus shown that indeed $z = 1$ and $z = \tau_1$ are two zeros of $Y_1(z) - 1$. Finally, there cannot exist another value, say z^*, such that $Y_1(z^*) = 1$. This is because, otherwise $K(z^*,1) = 0$ which is impossible since $K(z,1)$ has only two zeros, namely $z = 1$ and $z = \tau_1$. Since $z = \tau_1$ is the only root of $Y_1(z) - 1 = 0$ in the region $1 < |z| < \tau_T$, statement (1) is proven.

(2): From Lemma 2.4 we know that $|X_1(z)| < 1$ if $1 < |z| < \tau_2$. Substituting $\{z_1 = X(z), z_2 = z\}$ into (2.12) causes that the LHS vanishes, yielding

$$(1-\alpha)(z-1)X_1(z)U(X_1(z),0) = -\alpha(X_1(z)-1)zU(0,z), \quad 1 < |z| < \tau_2 \,. \tag{2.43}$$

According to Lemma 2.4, we have that $|X_1(z)| < \tau_1$ for $1 < |z| < \tau_T$. Consequently, $U(X_1(z),0)$ is still analytic inside this region. Hence, $\alpha(X_1(z) - 1)zU(0,z)$ can be analytically continued into the region $\tau_2 \leq |z| < \tau_T$ via (2.43). We conclude that the function $U(z,0)$ is meromorphic in $\tau_2 \leq |z| < \tau_T$, its poles being the zeros (if any) of $X_1(z) - 1$ in $\tau_2 \leq |z| < \tau_T$. Analogously as in in the proof of the first part of this theorem, it can be proven that $z = \tau_2$ is the only zero in this region. □

From Theorem 2.1 it follows that τ_1 is an isolated singularity of $U(z,0)$. Let us rewrite (2.42) as

$$U(z,0) = -\frac{\alpha(z-1)Y_1(z)U(0,Y_1(z))}{(1-\alpha)(Y_1(z)-1)z} \,.$$

Multiplying the equation above by $(z-\tau_1)$ and taking the limit to τ_1 yields

$$\lim_{z\to\tau_1}(z-\tau_1)U(z,0) = -\frac{\alpha-\lambda_1}{(1-\alpha)^2}\frac{1-\lambda_1-\lambda_2}{\lambda_1} \,. \tag{2.44}$$

Because the above limit is strictly negative (and hence different from zero), τ_1 is a simple pole of $U(z,0)$. Obviously, τ_1 must be the radius of convergence of the power series (2.24).

Analogously it follows that τ_2 is the radius of convergence of the power series (2.25) and that τ_2 is a simple pole of $U(0,z)$ with residue

$$\lim_{z\to\tau_2}(z-\tau_2)U(0,z) = -\frac{1-\alpha-\lambda_2}{\alpha^2}\frac{1-\lambda_1-\lambda_2}{\lambda_2}\,. \tag{2.45}$$

Continuation to $|z| \geq \tau_T$

The composition $Y_1(X_1(Y_1(z)))$

We start this section with a useful observation for values of z where $Y_1'(z)$ and $X_1'(z)$ vanish. Notice that we can compute $Y_1'(z)$ as

$$Y_1'(z) = -\frac{K^{(1)}(z,Y_1(z))}{K^{(2)}(z,Y_1(z))}\,. \tag{2.46}$$

Consider now a value z^* for which $Y_1'(z^*)=0$. From Equation (2.46) it follows that $K^{(1)}(z^*,Y_1(z^*))=0$. Hence, z^* is a zero with multiplicity (at least) two of the polynomial $z\mapsto K(z,Y_1(z^*))$. This is only possible if $D_X(Y_1(z^*))=0$. According to Lemma 2.3, we must have that $Y_1(z^*)$ is equal to either y_1, y_2, y_3 or y_4. Further, we cannot have that also $Y_1''(z^*)=0$, because otherwise $K^{(11)}(z^*,Y(z^*))=0$, which is impossible because $D_X(z)$ has no zeros with multiplicity three or more.

We now study the graph of $Y_1(z)$ on the real interval $[1,\tau_1]$. Note that $Y_1(z)$ and $Y_1'(z)$ are continuous on $[1,\tau_1]$. In the proof of Theorem 2.1, we have shown that $Y_1(1)=1$, $Y_1(\tau_1)=1$. Likewise, it can be shown that $Y_1(\tau_T)=\tau_2$. The derivative of $Y_1(z)$, evaluated at $z=1$, $z=\tau_1$ and $z=\tau_T$ is, using (2.46), given by

$$Y_1'(1) = -\frac{\alpha-\lambda_1}{1-\alpha-\lambda_2} < 0\,, \tag{2.47}$$

$$Y_1'(\tau_1) = \frac{(1-\alpha)\lambda_1(\alpha-\lambda_1)}{\alpha(1-\lambda_1)(1-\lambda_1-\lambda_2)} > 0\,, \tag{2.48}$$

and

$$Y_1'(\tau_T) = \frac{(1-\alpha)\lambda_1(1-\lambda_1-\lambda_2)}{\alpha(\alpha-\lambda_1)(1-\lambda_1)} > 0 \tag{2.49}$$

respectively. From (2.47) and (2.48) it follows that there exists a $v^*\in\,]1,\tau_1[$, such that $Y_1'(v^*)=0$. Because of Lemma 2.3 and the fact that $|Y_1(z)|<1$ for $1<|z|<\tau_1$, there can be at most two values such that the derivative of $Y_1'(z)$ in $]1,\tau_T[$ vanishes. Hence, v^* is the only value in $]1,\tau_1[$ such that $Y'(z)$ vanishes.
Similarly, there cannot exist a value in the interval $]\tau_1,\tau_T[$ such that $Y_1'(z)$ vanishes, since $|Y_1(z)|<\tau_2$ for $\tau_1\leq|z|<\tau_T$, $Y_1'(\tau_1)>0$, $Y_1'(\tau_T)>0$ and because of Lemma 2.3.
We conclude that $Y_1(z)$ is strictly decreasing for $z\in[1,v^*[$ and strictly increasing for $z\in\,]v^*,\tau_T]$.

Analogously, there exists a $w^* \in]1, \tau_2[$, such that $X_1(z)$ is strictly decreasing for $z \in [1, w^*[$ and strictly increasing for $z \in]w^*, \tau_T]$.

Without loss of generality, we assume that $\tau_1 \leq \tau_2$ in the remainder of this subsection. For every z, it holds that either $X_1(Y_1(z)) = z$ or $X_2(Y_1(z)) = z$, by definition of $X_1(z)$ and $X_2(z)$. For example, consider $z = 1$. Then we have that $X_1(Y_1(1)) = 1$, while $X_2(Y_1(1)) = \tau_1$. Starting in $z = 1$, we use the (real) inverse function theorem to investigate in what region the equation $X_1(Y_1(z)) = z$ holds. It can be seen that the inverse function theorem will work as long as $Y_1(z)$ is continuously differentiable and $Y_1'(z) \neq 0$. Because of the analysis above, this yields $X_1(Y_1(s)) = s$ if $s \in [1, v^*[$ and $X_2(Y_1(s)) = s$ if $s \in]v^*, \tau_T]$. From Equation (2.40), we get

$$X_1(Y_1(s)) = \tau_1 \frac{Y_1(s)}{s}, \quad s \in]v^*, \tau_T] \, .$$

Analogously, we have that

$$Y_1(X_1(s)) = \tau_2 \frac{X_1(s)}{s}, \quad s \in]w^*, \tau_T] \, .$$

Substituting $s = Y_1(s)$ into the above equation yields

$$Y_1(X_1(Y_1(s))) = \tau_2 \frac{X_1(Y_1(s))}{Y_1(s)}, \quad Y_1(s) \in]w^*, \tau_T] \, .$$

Because $Y_1(\tau_1) = 1$ and $Y_1(\tau_T) = \tau_2$, there exists a $t^* \in]\tau_1, \tau_T[$, such that $Y_1(t^*) = w^*$. Hence $Y_1(s) > w^*$ if $s \in]t^*, \tau_T]$, because $Y_1(s)$ is strictly increasing on $[t^*, \tau_T]$. Because also $v^* < \tau_1 \leq \tau_2 < t^*$, we have the following key property

$$Y_1(X_1(Y_1(s))) = \frac{\tau_T}{s}, \quad s \in]t^*, \tau_T] \, . \tag{2.50}$$

A new functional equation

From the proof in Theorem 2.1, we have that for $|z| < \tau_T$,

$$(1-\alpha)\frac{zU(z,0)}{z-1} = -\alpha \frac{Y_1(z)U(0, Y_1(z))}{Y_1(z) - 1} \tag{2.51}$$

$$\alpha \frac{zU(0,z)}{z-1} = -(1-\alpha)\frac{X_1(z)U(X_1(z), 0)}{X_1(z) - 1} \, . \tag{2.52}$$

If we restrict z to the real interval $]t^*, \tau_T[$ and use consecutively Equations (2.51), (2.52) and (2.50) we find that

$$\begin{aligned}(1-\alpha)\frac{zU(z,0)}{z-1} &= -\alpha \frac{Y_1(z)U(0, Y_1(z))}{Y_1(z) - 1} \\ &= (1-\alpha)\frac{X_1(Y_1(z))U(X_1(Y_1(z)), 0)}{X_1(Y_1(z)) - 1}\end{aligned}$$

$$= -\alpha \frac{Y_1(X_1(Y_1(z)))U(0, Y_1(X_1(Y_1(z))))}{Y_1(X_1(Y_1(z))) - 1}$$
$$= -\alpha\tau_T \frac{U(0, \frac{\tau_T}{z})}{\tau_T - z}, \quad z \in]t^*, \tau_T[\,, \tag{2.53}$$

which can be rewritten as

$$U(z,0) = -\frac{\alpha\tau_T(z-1)U(0, \frac{\tau_T}{z})}{(1-\alpha)z(\tau_T - z)}, \quad z \in]t^*, \tau_T[\,. \tag{2.54}$$

Since $U(0, \frac{\tau_T}{z})$ is analytic for $|z| > 1$, $z \neq \tau_1$, the right-hand side of (2.53) is analytic for $|z| > 1$, $z \neq \{\tau_1, \tau_T\}$. Hence, the equality is valid for $1 < |z| < \tau_T$, $z \neq \tau_1$ and $U(z,0)$ can be analytically continued for $|z| \geq \tau_T$, $z \neq \tau_T$ via Equation (2.54). $U(z,0)$ is therefore analytic in $\mathbb{C} \setminus \{\tau_1, \tau_T\}$. Using (2.54), it follows that

$$\lim_{z \to \tau_T} (z - \tau_T)U(z,0) = \frac{(\alpha - \lambda_1)(1 - \lambda_1 - \lambda_2)}{(1-\alpha)\lambda_1\lambda_2}\,, \tag{2.55}$$

hence τ_T is a simple pole of $U(z,0)$.

Further, using (2.53) it follows that $U(0,z)$ is analytic in $\mathbb{C} \setminus \{\tau_2, \tau_T\}$, and τ_T is a simple pole of $U(0,z)$ with residue

$$\lim_{z \to \tau_T} (z - \tau_T)U(0,z) = \frac{(1 - \alpha - \lambda_2)(1 - \lambda_1 - \lambda_2)}{\alpha\lambda_1\lambda_2}\,. \tag{2.56}$$

Using the new functional equation (2.54), it also easily follows that

$$\lim_{z \to \infty} U(z,0) = 0\,, \tag{2.57}$$
$$\lim_{z \to \infty} U(0,z) = 0\,. \tag{2.58}$$

We are now ready to present our main theorem.

Theorem 2.2.

1. *$U(z,0)$ is a rational function and its partial fraction expansion is given by*

$$U(z,0) = \frac{(\alpha - \lambda_1)(1 - \lambda_1 - \lambda_2)}{(1-\alpha)} \times \left(\frac{-1}{(1-\alpha)\lambda_1 z - \alpha(1-\lambda_1)} + \frac{1}{\lambda_1\lambda_2 z - (1-\lambda_1)(1-\lambda_2)} \right). \tag{2.59}$$

2. *$U(0,z)$ is a rational function and its partial fraction expansion is given by*

$$U(0,z) = \frac{(1 - \alpha - \lambda_2)(1 - \lambda_1 - \lambda_2)}{\alpha}$$

$$\times \left(\frac{-1}{\alpha\lambda_2 z - (1-\alpha)(1-\lambda_2)} + \frac{1}{\lambda_1\lambda_2 z - (1-\lambda_1)(1-\lambda_2)} \right). \tag{2.60}$$

Proof. We will only prove part 1, since part 2 can be obtained analogously.

$U(z,0)$ is an analytic function for all z with the exception of τ_1 and τ_T as simple poles. From (2.44) and (2.55), we obtain that the singular part of $U(z,0)$ at τ_1 and τ_T is given by $-\frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)}{(1-\alpha)^2\lambda_1(z-\tau_1)}$ and $\frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)}{(1-\alpha)\lambda_1\lambda_2(z-\tau_T)}$, respectively. Hence

$$L(z) := U(z,0) - \left(-\frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)}{(1-\alpha)^2\lambda_1(z-\tau_1)} + \frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)}{(1-\alpha)\lambda_1\lambda_2(z-\tau_T)} \right)$$

is an entire function. Because $U(z,0) \to 0$ as $z \to \infty$, we have that $L(z) \to 0$ as $z \to \infty$ as well. Hence $L(z)$ is bounded and tends to zero (as $z \to \infty$). From Liouville's theorem we can conclude that $L(z) = 0$. □

2.3.5 The joint distribution $p(n,m)$

Substituting Equations (2.59) and (2.60) into (2.16) yields

$$\begin{aligned} U(z_1,z_2) = {} & \frac{(\alpha-\lambda_1)(1-\alpha-\lambda_2)(1-\lambda_1-\lambda_2)(1-\lambda_1+\lambda_1 z_1)(1-\lambda_2+\lambda_2 z_2)}{(\alpha(1-\lambda_1)-(1-\alpha)\lambda_1 z_1)((1-\alpha)(1-\lambda_2)-\alpha\lambda_2 z_2)} \\ & \times \frac{(1-\lambda_1)(1-\lambda_2)-\lambda_1\lambda_2 z_1 z_2}{((1-\lambda_1)(1-\lambda_2)-\lambda_1\lambda_2 z_1)((1-\lambda_1)(1-\lambda_2)-\lambda_1\lambda_2 z_2)}\,. \end{aligned} \tag{2.61}$$

The joint PGF $U(z_1,z_2)$ is now completely determined in terms of the system parameters α, λ_1 and λ_2. From this joint PGF we can obtain the joint pmf.

Theorem 2.3. *The joint probability distribution $p(n,m)$ of the number of type-1 and type-2 customers is given by*

$$\begin{aligned} p(0,0) = {} & \frac{(1-\alpha-\lambda_2)\,(\alpha-\lambda_1)\,(1-\lambda_1-\lambda_2)}{(1-\lambda_2)\,\alpha\,(1-\alpha)\,(1-\lambda_1)} \\ p(n,0) = {} & \frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)}{(1-\alpha)(1-\lambda_1)} \left(\frac{1}{\alpha}\frac{1}{\tau_1^n} - \frac{1}{1-\lambda_2}\frac{1}{\tau_T^n} \right), \quad n \geq 0 \\ p(0,n) = {} & \frac{(1-\alpha-\lambda_2)(1-\lambda_1-\lambda_2)}{\alpha(1-\lambda_2)} \\ & \times \left(\frac{1}{(1-\alpha)}\frac{1}{\tau_2^n} - \frac{1}{1-\lambda_1}\frac{1}{\tau_T^n} \right), \quad n \geq 0 \\ p(n,m) = {} & \frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)}{\alpha(1-\alpha)\lambda_1(1-\lambda_1)}\frac{1}{\tau_1^n\tau_T^m} \\ & + \frac{(1-\alpha-\lambda_2)(1-\lambda_1-\lambda_2)}{\alpha(1-\alpha)\lambda_2(1-\lambda_2)}\frac{1}{\tau_2^m\tau_T^n} \\ & - \frac{(1-\lambda_1-\lambda_2)^2}{\lambda_1(1-\lambda_1)\lambda_2(1-\lambda_2)}\frac{1}{\tau_T^{n+m}}, \quad n \geq 1, m \geq 1\,. \end{aligned} \tag{2.62}$$

Proof. Expanding the factors in the denominator of (2.61) with respect to z_1 and z_2 yields the result. It can be verified that $\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\Pr[u_1 = n, u_2 = m] = 1$ and that $\Pr[u_1 = n, u_2 = m] \geq 0$, $\forall n, m \in \mathbb{N}$. □

2.3.6 The marginal distribution $p_T(n)$

Finally, from equation (2.61), we easily obtain an expression for the PGF $U_T(z)$ by substituting $z_1 = z, z_2 = z$,

$$U_T(z) = \frac{(\alpha-\lambda_1)(1-\alpha-\lambda_2)(1-\lambda_1-\lambda_2)(1-\lambda_1+\lambda_1 z)(1-\lambda_2+\lambda_2 z)}{(\alpha(1-\lambda_1)-(1-\alpha)\lambda_1 z)((1-\alpha)(1-\lambda_2)-\alpha\lambda_2 z)} \times \frac{(1-\lambda_1)(1-\lambda_2)-\lambda_1\lambda_2 z^2}{((1-\lambda_1)(1-\lambda_2)-\lambda_1\lambda_2 z)^2} \,. \tag{2.63}$$

This PGF describes the total number of customers in the system. Let us denote the pmf of the total number of customers in the system in steady state as $p_T(n)$, i.e.

$$p_T(n) \triangleq \lim_{k\to\infty} \Pr[u_{1,k} + u_{2,k} = n] \,. \tag{2.64}$$

Because

$$U_T(z) = \sum_{n=0}^{\infty} p_T(n) z^n \,,$$

the pmf $p_T(n)$ can be obtained from (2.63) by expanding the factors in the denominator and equating the coefficients of powers of z.

An alternative is to use the joint probility distribution from Theorem 2.3 and use the identity

$$p_T(n) = \sum_{k=0}^{n} p(k, n-k)$$

Both alternatives give rise to the following result.

Theorem 2.4. *The probability distribution of the total number of customers is given by*

$$\begin{aligned}
p_T(0) &= \frac{(1-\alpha-\lambda_2)\,(\alpha-\lambda_1)\,(1-\lambda_1-\lambda_2)}{(1-\lambda_2)\,\alpha\,(1-\alpha)\,(1-\lambda_1)} \\
p_T(n) &= \frac{(\alpha-\lambda_1)(1-\lambda_1-\lambda_2)(\alpha\lambda_2+(1-\alpha-\lambda_2)\lambda_1)}{(1-\alpha)\alpha\lambda_1(1-\lambda_1)(1-\alpha-\lambda_2)}\frac{1}{\tau_1^n} \\
&\quad + \frac{(1-\alpha-\lambda_2)(1-\lambda_1-\lambda_2)((1-\alpha)\lambda_1+(\alpha-\lambda_1)\lambda_2)}{(1-\alpha)\alpha\lambda_2(1-\lambda_2)(\alpha-\lambda_1)}\frac{1}{\tau_2^n} \\
&\quad - \frac{(1-\lambda_1-\lambda_2)((1-\lambda_1)(1-\lambda_1-\lambda_2)-\lambda_2(\alpha-\lambda_1))}{(1-\lambda_1)(1-\lambda_2)(\lambda_2(\alpha-\lambda_1)}\frac{1}{\tau_T^n}
\end{aligned}$$

$$- \frac{(1-\lambda_1-\lambda_2)((1-\lambda_2)(1-\lambda_1-\lambda_2)-\lambda_1(1-\alpha-\lambda_2))}{(1-\lambda_1)(1-\lambda_2)(\lambda_1(1-\alpha-\lambda_2)} \frac{1}{\tau_T^n}$$
$$- \frac{(1-\lambda_1-\lambda_2)^2}{\lambda_1(1-\lambda_1)\lambda_2(1-\lambda_2)} \frac{(n-1)}{\tau_T^n}, \quad n \geq 1 \,. \tag{2.65}$$

2.3.7 Calculation of numerical characteristics

Moments

The moments of the type-1 and type-2 system contents can be obtained from their PGFs, using the moment-generating property.

The mean type-1 system content is given by (using either (2.18), or (1.37) with $A''(1) = 0$)

$$\mathsf{E}[u_1] = \frac{\lambda_1(1-\lambda_1)}{\alpha-\lambda_1} \,. \tag{2.66}$$

The mean type-2 system content is given by

$$\mathsf{E}[u_2] = \frac{\lambda_2(1-\lambda_2)}{1-\alpha-\lambda_2} \,. \tag{2.67}$$

The mean total system content can be obtained using expression (2.63), yielding

$$\mathsf{E}[u_T] = \frac{(1-\lambda_1-\lambda_2)((1-\alpha)\lambda_1+\alpha\lambda_2-\lambda_1\lambda_2)}{(\alpha-\lambda_1)(1-\alpha-\lambda_2)} \tag{2.68}$$

It is easily verified that equations (2.66), (2.67) and (2.68) satisfy $\mathsf{E}[u_T] = \mathsf{E}[u_1] + \mathsf{E}[u_2]$ (which is expected since $u_T = u_1 + u_2$).

The variance of the type-1 system content is given by

$$\begin{aligned} \mathrm{var}[u_1] &= \left.\frac{d^2U_1(z)}{dz^2}\right|_{z=1} + \left.\frac{d^2U_1(z)}{dz^2}\right|_{z=1} - \left(\left.\frac{d^2U_1(z)}{dz^2}\right|_{z=1}\right)^2 \\ &= \frac{\lambda_1(1-\lambda_1)(\alpha+\lambda_1^2-2\alpha\lambda_1)}{(\alpha-\lambda_1)^2} \,. \end{aligned} \tag{2.69}$$

The variance of the type-2 system content is given by

$$\begin{aligned} \mathrm{var}[u_2] &= \left.\frac{d^2U_2(z)}{dz^2}\right|_{z=1} + \left.\frac{d^2U_2(z)}{dz^2}\right|_{z=1} - \left(\left.\frac{d^2U_2(z)}{dz^2}\right|_{z=1}\right)^2 \\ &= \frac{\lambda_2(1-\lambda_2)(1-\alpha+\lambda_2^2-2(1-\alpha)\lambda_2)}{(1-\alpha-\lambda_2)^2} \,. \end{aligned} \tag{2.70}$$

The so-called content covariance, i.e. the covariance between the type-1 and type-2 system contents at the same slot boundary is given by

$$\mathrm{cov}[u_1,u_2] = \frac{d^2U(z_1,z_2)}{dz_1dz_2} - \left(\left.\frac{dU_1(z)}{dz}\right|_{z=1}\right)\left(\left.\frac{dU_2(z)}{dz}\right|_{z=1}\right)$$

$$= -\frac{\lambda_1\lambda_2(1-\lambda_1)(1-\lambda_2)}{(1-\lambda_1-\lambda_2)^2} \ . \tag{2.71}$$

Expression (2.71) gives two noteworthy results. First, the covariance is independent of the system parameter α, which is quite remarkable. Notice that the correlation coefficient between both system contents does depend on the parameter α, because the variances of u_1 and u_2 depend on α. Secondly, the covariance is always negative, which could be more or less expected because if u_1 is exceptionally large, then either there were a lot of type-1 arrivals in the previous time slots, or type-1 customers were not served often during the last couple of slots. In the latter case, it is likely that u_2 is small.

Since $\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}[X,Y]$ for two random variables X and Y, we obtain the variance of the total system content

$$\begin{aligned}\mathrm{var}[u_T] = &\frac{\lambda_1(1-\lambda_1)(\alpha+\lambda_1^2-2\alpha\lambda_1)}{(\alpha-\lambda_1)} \\ &+ \frac{\lambda_2(1-\lambda_2)(1-\alpha+\lambda_2^2-2(1-\alpha)\lambda_2)}{(1-\alpha-\lambda_2)} - \frac{2\lambda_1\lambda_2(1-\lambda_1)(1-\lambda_2)}{(1-\lambda_1-\lambda_2)^2} \ .\end{aligned}$$

A measure of inefficiency

A performance measure related to the probability of an empty system $p(0,0)$ is the probability that the server selects an empty queue while the non-selected queue is non-empty. We emphasize that in case of a work-conserving scheduling discipline, this probability is necessarily zero. Let us define this probability as σ. For our model, we have

$$\begin{aligned}\sigma &= \lim_{k\to\infty} \left(\Pr[u_{1,k} > 0, u_{2,k} = 0, r_k = 0] + \Pr[u_{1,k} = 0, u_{2,k} > 0, r_k = 1]\right) \\ &= \lim_{k\to\infty} \left(\Pr[u_{1,k} > 0, u_{2,k} = 0](1-\alpha) + \Pr[u_{1,k} = 0, u_{2,k} > 0]\alpha\right) \\ &= \lim_{k\to\infty} \left(\Pr[u_{2,k} = 0](1-\alpha) + \Pr[u_{1,k} = 0]\alpha - \Pr[u_{1,k} = 0, u_{2,k} = 0]\right) \\ &= p_2(0)(1-\alpha) + p_1(0)\alpha - p(0,0) \ . \end{aligned} \tag{2.72}$$

Substituting the expressions for $p_1(0)$, $p_2(0)$ and $p(0,0)$ into the above, yields

$$\sigma = \frac{(1-\lambda_1-\lambda_2)((1-\alpha)^2\lambda_1 + \alpha^2\lambda_2 - (1-\alpha+\alpha^2)\lambda_1\lambda_2)}{\alpha(1-\alpha)(1-\lambda_1)(1-\lambda_2)} \ . \tag{2.73}$$

The mean maximum system content

As a final numerical characteristic, we compute the mean maximum system content $\mathsf{E}[\max(u_1,u_2)]$. The computation of this mean value is possible since we have an explicit expression for the joint pmf $p(n,m)$. Using the joint pmf (2.62)

of Theorem 2.3 and the fact that $\Pr[\max(u_1, u_2) > L] = 1 - \Pr[u_1 \leq L, u_2 \leq L]$, the expected maximum system content is given by

$$\begin{aligned}\mathsf{E}[\max(u_1, u_2)] &= \sum_{L=0}^{\infty} \Pr[\max(u_1, u_2) > L] \\ &= \frac{\tau_T}{\tau_T^2 - 1} + \frac{\lambda_1(1-\lambda_1)}{\alpha - \lambda_1} - \frac{(1-\lambda_1)(1-\lambda_2)}{\alpha(1-\lambda_1)\tau_T - (1-\alpha)\lambda_1} \\ &\quad + \frac{\lambda_2(1-\lambda_2)}{1-\alpha-\lambda_2} - \frac{(1-\lambda_1)(1-\lambda_2)}{(1-\alpha)(1-\lambda_2)\tau_T - \alpha\lambda_2} \,. \end{aligned} \tag{2.74}$$

2.3.8 Some numerical examples

In this section we present some numerical examples to show the influence of the system parameters on the performance measures. We focus on the measure of inefficiency σ and the mean maximum system content $\mathsf{E}[\max(u_1, u_1)]$.

Figure 2.1 shows the measure of inefficiency σ versus the (scaled) mean type-2 arrival rate with $\lambda_1 = 0.3$ and $\alpha = 0.31$, 0.5, 0.55 and 0.69 respectively. Because the stability condition requires that $\lambda_2 < 1 - \alpha$, we have scaled the horizontal axis by dividing by $1 - \alpha$, such that the four curves have the same domain $[0, 1[$. Remark that σ is a measure of **in**efficiency, with values between 0 and 1. Values close to 0 correspond to an efficient system, while values close to 1 correspond to an inefficient system. We say that the server is allocated to the wrong queue, if it is allocated to an empty queue while the other queue is non-empty. By definition (2.73) of σ, the more slots that the server is allocated to the wrong queue, the higher the value of σ.

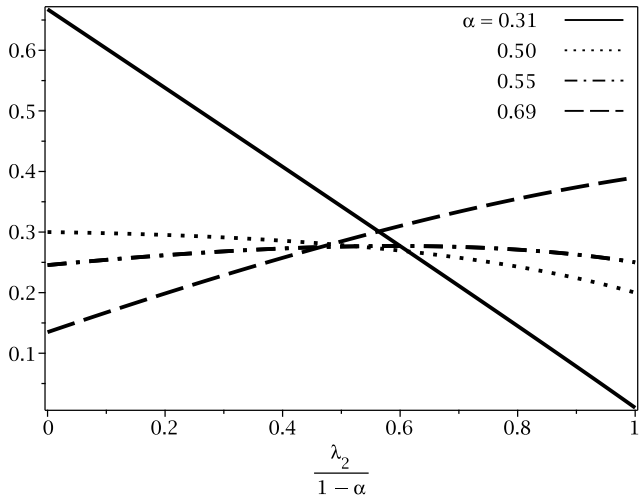


Figure 2.1: Measure of inefficiency σ versus the type-2 arrival rate ($\lambda_1 = 0.3$ fixed).

Let us start with the simplest case, namely when $\alpha = 0.5$. In this case, the two different queues get an equal share of the single-server capacity. If we increase λ_2, more type-2 customers arrive, leading to a higher type-2 system content (on average). Therefore, it is expected that if we increase λ_2, the server idles less when it is allocated to queue-2. Hence, the queueing system works more efficiently because there are less slots when the server is allocated to the wrong queue. This is in accordance with Figure 2.1 . We observe that σ decreases with increasing λ_2, at least for $\alpha = 0.5$. Looking at Figure 2.1, we see different behaviors in case of asymmetric weights. Let us assume that $0 < \alpha < 0.5$. Recall that we still assume that $\lambda_1 = 0.3$ is fixed. Because the stability condition for queue-1 has to be satisfied, we actually assume $0.3 < \alpha < 0.5$. In this case, more capacity of the server is distributed to queue-2. Hence, if we increase λ_2 (and thus the type-2 system content) this capacity is more often used. Looking at Figure 2.1 with $\alpha = 0.31$ (solid line), we see indeed a greater efficiency of the system for increasing λ_2. Finally, let us consider the case of $0.5 < \alpha < 1$. Discussing the (in)efficiency in this case is trickier. Remark that in this case most capacity of the server is distributed to the first queue. For ease of explanation, let us assume that $\alpha = 0.69$. Hence, most of the time the server is willing to serve a type-1 customer. However, since $\lambda_1 = 0.3$, the server will be idle during a lot of slots, regardless of the number of type-2 customers. Because there are many slots that the server idles, we have that the more type-2 customers arrive, the more slots the server is allocated to the wrong queue. Hence, in this case the system becomes less efficient for increasing λ_2. The same reasoning explains why in Figure 2.1 with $\alpha = 0.55$, σ increases for increasing λ_2 until a certain λ_2-value.

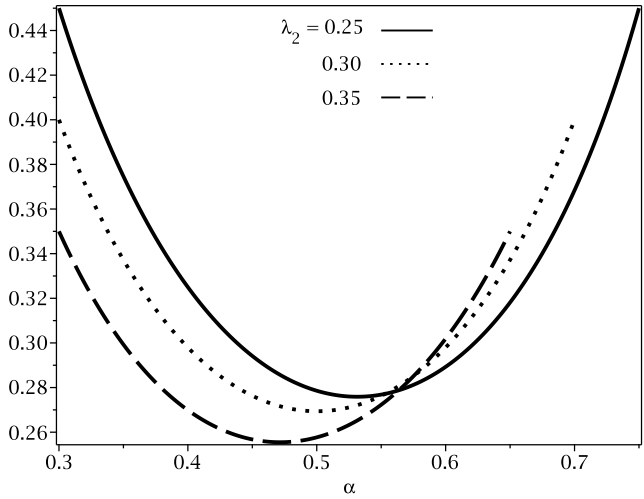


Figure 2.2: Measure of inefficiency σ versus α ($\lambda_1 = 0.3$ fixed).

We look again at σ, but with varying α while keeping the mean type-2 arrival

rate fixed. Figure 2.2 shows the measure of inefficiency σ versus α with $\lambda_1 = 0.3$ and $\lambda_2 = 0.2$, 0.3 and 0.4. The stability condition requires $\lambda_1 < \alpha < 1 - \lambda_2$. In contrast to Figure 2.1, we did not scale the horizontal axis in Figure 2.2. From Figure 2.2, we see that σ first decreases and then increases for increasing α. This can be explained as follows. If α increases, more server capacity is distributed to type-1 customers and less server capacity is distributed to type-2 customers. This trade-off results in a more efficient use of the server, starting from α close to $\lambda_1 = 0.3$. This positive effect continues until α achieves a critical value and from there, the negative effect of capacity loss for type-2 customers dominates the positive effect of capacity gain for type-1 customers.

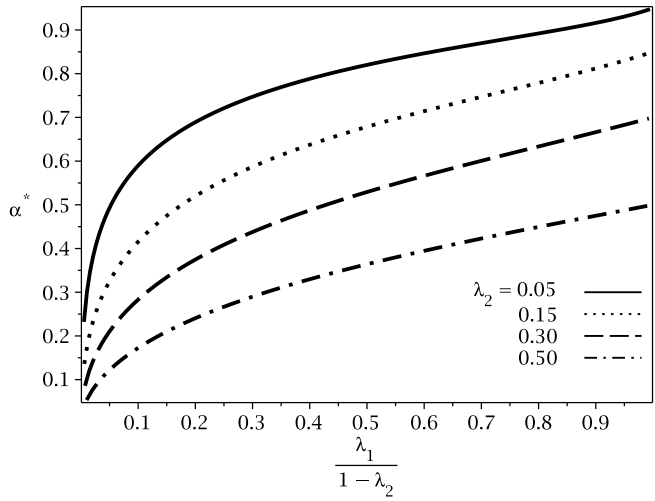


Figure 2.3: The optimal value α^* that minimizes the expected maximum system content $\mathsf{E}[\max(u_1, u_2)]$ versus the type-1 arrival rate λ_1, for various values of λ_2.

In practice of designing concrete systems, it is natural to determine the parameter α such that a certain cost function is minimized. Consider for example the expected maximum system content $\mathsf{E}[\max(u_1, u_2)]$ as the cost function for our queueing system, in which α can be considered as the decision variable. We want to compute the optimal value of α, denoted by α^*, that minimizes the cost function, i.e. we want to compute

$$\alpha^* = \arg\min_{\alpha} \mathsf{E}[\max(u_1, u_2)], \quad \lambda_1 < \alpha < 1 - \lambda_2 \,. \tag{2.75}$$

$\mathsf{E}[\max(u_1, u_2)]$ is, as a function in α, a rational function with poles λ_1, $1 - \lambda_2$, $\frac{\lambda_1}{(1-\lambda_1)\tau_T + \lambda_1}$ and $\frac{(1-\lambda_2)\tau_T}{\lambda_2 + (1-\lambda_2)\tau_T}$. Using the fact that $\tau_T > 1$, it follows that $\frac{\lambda_1}{(1-\lambda_1)\tau_T + \lambda_1} < \lambda_1$ and $1 - \lambda_2 < \frac{(1-\lambda_2)\tau_T}{\lambda_2 + (1-\lambda_2)\tau_T}$ such that $\mathsf{E}[\max(u_1, u_2)]$ is well defined for $\alpha \in]\lambda_1, 1 - \lambda_2[$. Moreover, $\mathsf{E}[\max(u_1, u_2)]$ is strictly positive for $\alpha \in]\lambda_1, 1 - \lambda_2[$ by definition. The optimization problem (2.75) can easily be solved using numerical software. Fig. 2.3 illustrates the optimal values of α

as a function of λ_1, for $\lambda_2 = 0.05$, 0.15, 0.3 and 0.5. Because the stability conditions imply that $\lambda_1 + \lambda_2 < 1$, we scaled the horizontal axis by dividing by $1 - \lambda_2$, such that four curves have the same domain $[0, 1]$. From Fig. 2.3, it can be seen that the optimal α, as a function of λ_1, increases the most for λ_1 close to 0. When $\lambda_1 \to 1 - \lambda_2$, the optimal α approaches $1 - \lambda_2$. This is due to the stability conditions $\lambda_1 < \alpha < 1 - \lambda_2$.

2.4 Identical Bernoulli arrivals in the two queues

As a second specific arrival distribution, suppose that during any slot k, the number of type-1 arrivals is the same as the number of type-2 arrivals, i.e.

$$a_{1,k} = a_{2,k}, \quad \text{for all } k \, .$$

We introduce the following new notation for the mean arrival rates,

$$\lambda_1 = \lambda_2 \triangleq \lambda \, . \tag{2.76}$$

Moreover, we assume that for given j the random variables $a_{j,k}$ constitute a sequence of independent and identically Bernoulli distributed random variables. In this case, the joint PGF $A(z_1, z_2)$ of the arrival process can be written as

$$A(z_1, z_2) = 1 - \lambda + \lambda z_1 z_2 \, . \tag{2.77}$$

This is the simplest model in the case that the number of arrivals of type-1 and type-2 customers are identical, i.e. $A(z_1, z_2) = C(z_1 z_2)$ with $C(z)$ a PGF. In contrast to the previous section, we now have that the numbers of type-1 and type-2 arrivals are correlated. Indeed, the covariance of $a_{1,k}$ and $a_{2,k}$ is given by

$$\text{cov}[a_{1,k}, a_{2,k}] = \lambda(1 - \lambda) \, , \tag{2.78}$$

which is always positive.

The mathematical analysis of the functional equation (2.12) with $A(z_1, z_2)$ given by (2.77) turns out to be considerably easier as compared to the analysis in Section 2.3.

2.4.1 The marginal distributions $p_1(n)$ and $p_2(n)$

Because the marginal arrivals are Bernoulli distributed, the expressions in Section 2.3.1 are also applicable here. For definiteness, we have that

$$\begin{aligned} p_1(0) &= 1 - \frac{\lambda}{\alpha} \, , \\ p_1(n) &= \frac{\alpha - \lambda}{\alpha(1 - \alpha)} \frac{1}{\tau_1^n}, \quad n \geq 1 \, , \end{aligned}$$

$$p_2(0) = 1 - \frac{\lambda}{1-\alpha} ,$$
$$p_2(n) = \frac{1-\alpha-\lambda}{\alpha(1-\alpha)} \frac{1}{\tau_2^n}, \quad n \geq 1 .$$

The values τ_1 and τ_2 are given by

$$\tau_1 = \frac{\alpha}{1-\alpha} \frac{1-\lambda}{\lambda} , \tag{2.79}$$
$$\tau_2 = \frac{1-\alpha}{\alpha} \frac{1-\lambda}{\lambda} . \tag{2.80}$$

As before, we have that $K(1,1) = 0$, $K(\tau_1, 1) = 0$ and $K(1, \tau_2) = 0$.

Finally, we consider the total system content. The corresponding PGF $U_T(z)$ is given by

$$\begin{aligned} U_T(z) &\triangleq U(z,z) \\ &= \frac{1-\lambda+\lambda z^2}{1-\lambda-\lambda z}[(1-\alpha)U(z,0) + \alpha U(0,z)]. \end{aligned} \tag{2.81}$$

This expression follows from Equation (2.12) with $z_1 = z_2 = z$ and in which we canceled the common factor $z(z-1)$ in numerator and denominator. The dominant singularity of $U_T(z)$ is either the dominant singularity of $U(z,0)$, the dominant singularity of $U(0,z)$, or the zero of the denominator. Note that the latter is given by $\frac{1-\lambda}{\lambda}$.

2.4.2 Areas of convergence

It is not surprising that the same conclusions as in Section 2.3.2 also apply here, since the results of Section 2.3.2 are based on the marginal distributions $p_1(n)$ and $p_2(n)$. Consequently, $U(z_1, z_2)$ is analytic in the two polydiscs $|z_1| < \tau_1$, $|z_2| \leq 1$ and $|z_1| \leq 1$, $|z_2| < \tau_2$. Furthermore, the partial PGFs $U(z_1, 0)$ and $U(0, z_2)$ are analytic in $|z_1| < \tau_1$ and $|z_2| < \tau_2$, respectively.

2.4.3 Analysis of the kernel $K(z_1, z_2)$

Substituting (2.77) into (2.13) yields

$$K(z_1,z_2) = z_1 z_2 - (1-\alpha)(1-\lambda)z_1 - (1-\alpha)\lambda z_1^2 z_2 - \alpha(1-\lambda)z_2 - \alpha\lambda z_1 z_2^2 . \tag{2.82}$$

The kernel $K(z_1, z_2)$ has the same form as the kernel that is studied in [81]. We follow the same steps as in [81] to analyze the zeros of the kernel K. To that end, let us define

$$H_1(z_1) = (1-\alpha)\lambda + \alpha(1-\lambda) - (1-\alpha)\lambda z_1 - \alpha(1-\lambda)\frac{1}{z_1} \tag{2.83}$$

$$H_2(z_2) = (1-\alpha)(1-\lambda) + \alpha\lambda - \alpha\lambda z_2 - (1-\alpha)(1-\lambda)\frac{1}{z_2} , \tag{2.84}$$

such that the kernel K can be rewritten as

$$K(z_1, z_2) = z_1 z_2 (H_1(z_1) + H_2(z_2)) .$$

We have the convenient property that for complex values of $z_1 = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}} e^{\mathrm{i}\theta}$, $H_1(z_1)$ is real and equal to

$$(1-\alpha)\lambda + \alpha(1-\lambda) - 2\sqrt{\alpha(1-\alpha)\lambda(1-\lambda)}\cos(\theta) ,$$

which can be easily established by using the well-known identity $e^{\mathrm{i}\theta} + e^{-\mathrm{i}\theta} = 2\cos(\theta)$. Further, we have that

$$(1-\alpha)\lambda + \alpha(1-\lambda) - 2\sqrt{\alpha(1-\alpha)\lambda(1-\lambda)}\cos\theta > 0 , \tag{2.85}$$

because

$$\begin{aligned}
&(1-\alpha)\lambda + \alpha(1-\lambda) - 2\sqrt{\alpha(1-\alpha)\lambda(1-\lambda)}\cos\theta \\
&\geq (1-\alpha)\lambda + \alpha(1-\lambda) - 2\sqrt{\alpha(1-\alpha)\lambda(1-\lambda)} \\
&= (\sqrt{(1-\alpha)\lambda} - \sqrt{\alpha(1-\lambda)})^2 \\
&> 0 .
\end{aligned}$$

We are now ready to prove the following lemma.

Lemma 2.5. *For values $z_1 = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}} e^{\mathrm{i}\theta}$, there is a unique $z_2 =: y(z_1) \in]0,1[$ such that*

$$H_1(z_1) + H_2(y(z_1)) = 0.$$

Proof. For $x \in [0,1]$, the function $H_2(x)$ increases monotonically from $-\infty$ at $x = 0$ to 0 at $x = 1$. Moreover, it holds that,

$$H_1\left(\sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}} e^{\mathrm{i}\theta}\right) > 0 ,$$

cf. (2.85). By virtue of the intermediate value theorem, there is a unique value z_2 in the interval $]0,1[$ such that

$$H_2(z_2) = -H_1\left(\sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}} e^{\mathrm{i}\theta}\right) .$$

□

2.4.4 Analytic continuation of $U(z,0)$ and $U(0,z)$

We can now proceed to determine the functions $U(z,0)$ and $U(0,z)$ and hence solve the functional equation (2.12). To accomplish this, it is crucial to note that

$$y(z) = y(\bar{z}), \quad |z| = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}},$$

with y defined in the previous subsection. The reason why the equality above holds, is simply because

$$\begin{aligned} H_1\left(\sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}}e^{\mathrm{i}\theta}\right) &= (1-\alpha)\lambda + \alpha(1-\lambda) - 2\sqrt{\alpha(1-\alpha)\lambda(1-\lambda)}\cos(\theta) \\ &= (1-\alpha)\lambda + \alpha(1-\lambda) - 2\sqrt{\alpha(1-\alpha)\lambda(1-\lambda)}\cos(-\theta) \\ &= H_1\left(\sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}}e^{-\mathrm{i}\theta}\right) . \end{aligned}$$

Further, since $\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} > 1$ it obviously holds that

$$\sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}} < \frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} = \tau_1 .$$

Due to the inequality above and the fact that $y(z) \in]0,1[$, we have that $U(z,y(z))$ remains finite. Hence, substituting $\{z_1 = z, z_2 = y(z)\}$, $|z| = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}}$ into the functional equation (2.12) yields

$$(1-\alpha)(y(z)-1)zU(z,0) + \alpha(z-1)y(z)U(0,y(z)) = 0 ,$$

and substituting $\{z_1 = \bar{z}, z_2 = y(z)\}$, $|z| = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}}$ into the functional equation (2.12) yields

$$(1-\alpha)(y(z)-1)\bar{z}U(\bar{z},0) + \alpha(\bar{z}-1)y(z)U(0,y(z)) = 0 .$$

Eliminating $U(0,y(z))$ gives us

$$(\bar{z}-1)zU(z,0) = (z-1)\bar{z}U(\bar{z},0) .$$

When multiplying both sides of the relation above by z and using the relations $z\bar{z} = \frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} \Leftrightarrow |z| = |\bar{z}| = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}}$, we immediately get

$$z\left(\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} - z\right)U(z,0) = \frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}(z-1)U\left(\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}z^{-1}, 0\right) , \quad (2.86)$$

with $|z| = \sqrt{\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}}$.

We observe that the LHS of (2.86) is analytic in $|z| < \tau_1$. Due to

$$\left| \frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} z^{-1} \right| = \tau_1 |z|^{-1}$$

and the fact that $U(z,0)$ is analytic in $|z| < \tau_1$, it follows that the RHS of (2.86) is analytic in $|z| > 1$. Both sides of (2.86) have a common region of analyticity, namely the region $1 < |z| < \tau_1$. Hence, using analytic continuation we conclude that both sides of (2.86), in their respective regions of analyticity, are equal to the same unique entire function, say $h(z)$. The LHS of (2.86) is clearly bounded for $|z| < \tau_1$. Moreover, the RHS of (2.86) is bounded by a first degree polynomial for $|z| > 1$, because

$$\lim_{|z|\to+\infty} U\left(\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} z^{-1}, 0 \right) = U(0,0) = p(0,0) < 1 \; .$$

Hence, the entire function $h(z)$ must be a first degree polynomial by virtue of Liouville's theorem. Hence, we have obtained that for $|z| < \tau_1$,

$$z\left(\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda} - z \right) U(z,0) = h(z) = C_1 z + C_2$$

with C_j to be determined. Substituting $z = 0$ immediately yields that $C_2 = 0$. Substituting $z_1 = 1$ and using that $U(1,0) = p_2(0) = 1 - \frac{\lambda}{1-\alpha}$, gives us

$$C_1 = \frac{\alpha - \lambda}{(1-\alpha)\lambda} \left(1 - \frac{\lambda}{1-\alpha} \right) \; .$$

We have thus obtained the following expression for $U(z,0)$:

$$U(z,0) = \frac{\alpha - \lambda}{\alpha(1-\lambda) - (1-\alpha)\lambda z} \left(1 - \frac{\lambda}{1-\alpha} \right) \; . \tag{2.87}$$

For reasons of symmetry, we have that $U(0,z)$ is given by

$$U(0,z) = \frac{1 - \alpha - \lambda}{(1-\alpha)(1-\lambda) - \alpha\lambda z} \left(1 - \frac{\lambda}{\alpha} \right) \; . \tag{2.88}$$

Alternative solution method for equation (2.86)

If we substitute $U(z,0) = \sum p(n,0) z^n$ into (2.86), we obtain that

$$z(\tau_1 - z) \sum_{n=0}^{\infty} p(n,0) z^n = \tau_1 (z-1) \sum_{n=0}^{\infty} p(n,0) \tau_1^n z^{-n} \; ,$$

where we wrote τ_1 instead of $\frac{\alpha(1-\lambda)}{(1-\alpha)\lambda}$ for ease of notation. The equation above can be rewritten as

$$\sum_{n=0}^{\infty} \tau_1 p(n,0) z^{n+1} - \sum_{n=0}^{\infty} p(n,0) z^{n+2} = \sum_{n=0}^{\infty} p(n,0) \tau_1^{n+1} z^{-n+1} - \sum_{n=0}^{\infty} p(n,0) \tau_1^{n+1} z^{-n} , \quad (2.89)$$

or

$$\sum_{n=1}^{\infty} \tau_1 p(n-1,0) z^n - \sum_{n=2}^{\infty} p(n-2,0) z^n - \sum_{n=0}^{\infty} p(n+1,0) \tau_1^{n+2} z^{-n} - p(0,0)\tau_1 z + \sum_{n=0}^{\infty} p(n,0) \tau_1^{n+1} z^{-n} = 0 .$$

This equation is valid for z-values such that $|z| = \sqrt{\tau_1}$. Hence, by multiplying by appropriate powers of z and integrating over the positively oriented circle centered at 0 with radius $|z| = \sqrt{\tau_1}$ we obtain the following relations

$$\begin{aligned} p(0,0) - p(1,0)\tau_1 &= 0 \\ \tau_1 p(n,0) - p(n-1,0) &= 0, \ n \geq 1 . \end{aligned}$$

Solving this elementary difference equation yields

$$p(n,0) = \frac{1}{\tau_1^n} p(0,0) .$$

Finally, using the condition $\sum_{n=0}^{\infty} p(n,0) = p_2(0) = 1 - \frac{\lambda}{1-\alpha}$ gives us

$$p(0,0) = \frac{(\alpha - \lambda)(1 - \alpha - \lambda)}{\alpha(1-\alpha)(1-\lambda)} .$$

Hence, the sequence $p(n,0)$ is completely determined. The obtained expressions for $p(n,0)$ and $p(0,0)$ lead to the same result for $U(z,0)$. It is worth noting that $\frac{p(n,0)}{p_2(0)}$ is a geometric distribution with parameter $1 - \frac{1}{\tau_1} = \frac{\alpha - \lambda}{\alpha(1-\lambda)}$. The sequence $\frac{p(n,0)}{p_2(0)}$ corresponds to the conditional probability distribution of the type-1 system content, given that the type-2 system content is zero. Analogously, it can be shown that $\frac{p(0,n)}{p_1(0)}$ is a geometric distribution with parameter $1 - \frac{1}{\tau_2} = \frac{1-\alpha-\lambda}{(1-\alpha)(1-\lambda)}$.

2.4.5 The joint distribution $p(n,m)$

Substituting (2.87) and (2.88) into (2.12) yields

$$U(z_1, z_2) = \frac{(1 - \lambda + \lambda z_1 z_2)(\alpha - \lambda)(1 - \alpha - \lambda)}{(\alpha(1-\lambda) - (1-\alpha)\lambda z_1)((1-\alpha)(1-\lambda) - \alpha\lambda z_2)} . \quad (2.90)$$

The joint PGF $U(z_1, z_2)$ is now completely determined in terms of the system parameters λ and α. From this PGF we can obtain the joint pmf.

Theorem 2.5. *The joint probability mass function of type-1 and type-2 system contents is given by*

$$\begin{aligned} p(0,0) &= \frac{(\alpha-\lambda)(1-\alpha-\lambda)}{\alpha(1-\alpha)(1-\lambda)} \,, \\ p(n,0) &= \frac{(\alpha-\lambda)(1-\alpha-\lambda)}{\alpha(1-\alpha)(1-\lambda)} \frac{1}{\tau_1^n}, \quad n \geq 0 \,, \\ p(0,n) &= \frac{(\alpha-\lambda)(1-\alpha-\lambda)}{\alpha(1-\alpha)(1-\lambda)} \frac{1}{\tau_2^n}, \quad n \geq 0 \,, \\ p(n,m) &= \frac{(\alpha-\lambda)(1-\alpha-\lambda)}{\alpha(1-\alpha)(1-\lambda)\lambda} \frac{1}{\tau_1^n} \frac{1}{\tau_2^m}, \quad n \geq 1, m \geq 1 \,. \end{aligned} \tag{2.91}$$

Proof. Expanding the factors in the denominator of (2.90) with respect to z_1 and z_2 gives the result. □

Looking more closely at expression (2.90), we observe that

$$U(z_1,z_2) = A(z_1,z_2) \frac{U_1(z_1)}{A_1(z_1)} \frac{U_2(z_2)}{A_2(z_2)} \,.$$

Let us rewrite this equation as follows

$$\frac{U(z_1,z_2)}{A(z_1,z_2)} = \frac{U_1(z_1)}{A_1(z_1)} \frac{U_2(z_2)}{A_2(z_2)} \,. \tag{2.92}$$

The equation above might ring a bell to some. Actually, the LHS is nothing else than the joint PGF of the *queue contents*, defined as the number of customers in the queue (thus without the one in the server if any). The two functions in the RHS are the PGFs of the marginal queue contents. The queue contents can be easily derived from the system contents. For the sake of exposition, let us define in this subsection $q_{1,k}$ and $q_{2,k}$ as the type-1 and type-2 queue content at the beginning of the k-th slot. We then get the following relation between the $q_{j,k}$ and the $u_{j,k}$:

$$u_{j,k+1} = q_{j,k} + a_{j,k}, \quad j = 1,2 \,. \tag{2.93}$$

This relation is explained as follows: the system content at the beginning of slot $k+1$ consist of the queue content at the beginning of the previous slot (the possible customer in the server during slot k has left the system at the end of that slot) and the customers that arrive during slot k. Let us denote the joint PGF of the stationary queue contents by $Q(z_1,z_2)$, i.e.

$$Q(z_1,z_2) = \lim_{k\to\infty} \mathsf{E}[z_1^{q_{1,k}} z_2^{q_{2,k}}] \,.$$

From equations (2.93) we find

$$U(z_1,z_2) = Q(z_1,z_2) A(z_1,z_2) \,.$$

From equation (2.92), we conclude that

$$Q(z_1, z_2) = Q(z_1, 1)Q(1, z_2) \; .$$

Since the joint PGF can be written as the product of the two marginal PGFs, this proves that the two queue contents are statistically independent.

2.4.6 The marginal distribution $p_T(n)$

From the two-dimensional PGF $U(z_1, z_2)$, we can derive an expression for the PGF $U_T(z)$ of the total system content at the beginning of an arbitrary slot, yielding

$$\begin{aligned} U_T(z) &= U(z, z) \\ &= \frac{(1 - \lambda + \lambda z^2)(\alpha - \lambda)(1 - \alpha - \lambda)}{(\alpha(1 - \lambda) - (1 - \alpha)\lambda z)((1 - \alpha)(1 - \lambda) - \alpha\lambda z)} \end{aligned} \tag{2.94}$$

The pmf of the total system content can then be obtained by computing the Taylor series expansion of $U_T(z)$ at $z = 0$ and identifying the coefficient at z^n with $p_T(n)$. We obtain the following result:

$$\begin{aligned} p_T(0) &= \frac{(\alpha - \lambda)(1 - \alpha - \lambda)}{\alpha(1 - \alpha)(1 - \lambda)} \; , \\ p_T(1) &= \frac{\lambda(\alpha - \lambda)(1 - \alpha - \lambda)(1 - 2\alpha + 2\alpha^2)}{\alpha^2(1 - \alpha)^2(1 - \lambda)^2} \; , \\ p_T(n) &= \frac{(\alpha - \lambda)(1 - \alpha - \lambda)}{\alpha(1 - \alpha)(1 - 2\alpha)\lambda(1 - \lambda)} \\ &\quad \times \left(\left(\lambda(1 - 2\alpha) + \alpha^2\right) \frac{1}{\tau_1^n} + \left(\lambda(1 - 2\alpha) - (1 - \alpha)^2\right) \frac{1}{\tau_2^n} \right) , \; n \geq 2 \; . \end{aligned}$$

2.4.7 Calculation of numerical characteristics

Moments

In this subsection, we give the expressions for the mean and the variance of the type-1, type-2 and total system contents.

The mean type-1 and type-2 system contents are given by, cf. (2.66),

$$\mathsf{E}[u_1] = \frac{\lambda(1 - \lambda)}{\alpha - \lambda} \; , \tag{2.95}$$

$$\mathsf{E}[u_2] = \frac{\lambda(1 - \lambda)}{1 - \alpha - \lambda} \; . \tag{2.96}$$

Furthermore, the mean total system content is given by, cf. (2.68),

$$\mathsf{E}[u_T] = \frac{(1 - 2\lambda)\lambda(1 - \lambda)}{(\alpha - \lambda)(1 - \alpha - \lambda)} \; . \tag{2.97}$$

Likewise, the variances of type-1 and type-2 system contents are given by, cf. (2.69),

$$\mathrm{var}[u_1] = \frac{\lambda(1-\lambda)(\alpha+\lambda^2-2\alpha\lambda)}{\alpha-\lambda} , \tag{2.98}$$

$$\mathrm{var}[u_2] = \frac{\lambda(1-\lambda)(1-\alpha+\lambda^2-2(1-\alpha)\lambda)}{1-\alpha-\lambda} . \tag{2.99}$$

The content covariance is found as

$$\begin{aligned}\mathrm{cov}[u_1,u_2] &= \frac{d^2U(z_1,z_2)}{dz_1dz_2} - \left(\left.\frac{dU_1(z)}{dz}\right|_{z=1}\right)\left(\left.\frac{dU_2(z)}{dz}\right|_{z=1}\right) \\ &= \lambda(1-\lambda) . \end{aligned} \tag{2.100}$$

Note that this expression is always positive. In fact, this is the same expression as the covariance of $a_{1,k}$ and $a_{2,k}$, cf. (2.78). However, we remark that the arrival correlation $\mathrm{corr}[a_{1,k},a_{2,k}]$ and the content correlation $\mathrm{corr}[u_1,u_2]$ are not equal to each other.

Finally, the variance of the total system content can be calculated by taking the appropriate derivatives of $U_T(z)$. Alternatively, the variance can also be obtained using the formula for the variance of the sum of two random variables. In either way, we obtain that

$$\begin{aligned}\mathrm{var}[u_T] = &\frac{\lambda(1-\lambda)(\alpha+\lambda^2-2\alpha\lambda)}{\alpha-\lambda} \\ &+ \frac{\lambda(1-\lambda)(1-\alpha+\lambda^2-2(1-\alpha)\lambda)}{1-\alpha-\lambda} + 2\lambda(1-\lambda) . \end{aligned} \tag{2.101}$$

A measure of inefficiency

Again, the probability σ that the server selects an empty queue while the non-selected queue is non-empty is given by

$$\sigma = p_2(0)(1-\alpha) + p_1(0)\alpha - p(0,0) .$$

For the derivation of the formula above, we refer to Section 2.3.7. In the current section, the performance measure σ is given by

$$\sigma = \frac{\lambda(1-\lambda-3\alpha+2\alpha\lambda+3\alpha^2-2\alpha^2\lambda)}{\alpha(1-\alpha)(1-\lambda)} . \tag{2.102}$$

The mean maximum system content

The mean maximum system content is found by first computing

$$\lim_{k\to\infty} \Pr[\max(u_{1,k},u_{2,k}) > L]$$

from the joint pmf (2.91). This can for instance be done by writing

$$\lim_{k\to\infty} \Pr[\max(u_{1,k}, u_{2,k}) > L] = 1 - \lim_{k\to\infty} \Pr[u_{1,k} \le L, u_{2,k} \le L] \; .$$

Secondly, using the well-known formula for the mean of a discrete random variable X: $\mathsf{E}[X] = \sum_{k=0}^{\infty} \Pr[X > k]$, the mean maximum system content is obtained as

$$\begin{aligned} \mathsf{E}[\max(u_1, u_2)] &= \sum_{L=0}^{\infty} \lim_{k\to\infty} \Pr[\max(u_{1,k}, u_{2,k}) > L] \\ &= \frac{(1-\lambda)\lambda}{1-\alpha-\lambda} + \frac{(1-\lambda)\lambda}{\alpha-\lambda} - \frac{\lambda(1-\lambda)}{1-2\lambda} \; . \end{aligned} \tag{2.103}$$

2.4.8 Some numerical examples

To conclude this section, we demonstrate the influence of the system parameters on the measure of inefficiency σ and the mean maximum system content $\mathsf{E}[\max(u_1, u_1)]$.

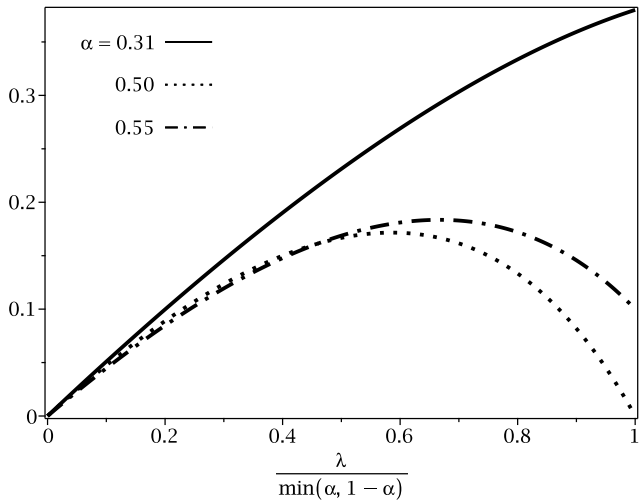


Figure 2.4: Measure of inefficiency σ versus the scaled arrival rate.

Figure 2.4 shows the measure of inefficiency σ versus the scaled arrival rate with $\alpha = 0.31,\ 0.5,\ 0.55$. Because the stability condition requires $\lambda < \alpha$ and $\lambda < 1-\alpha$, we have scaled the horizontal axis by dividing by $\min(\alpha, 1-\alpha)$, such that the three curves have the same domain $[0, 1[$. Obviously, for $\lambda = 0$, there are no customers in the system and hence no slots are wasted ($\sigma = 0$). It is interesting that for $\alpha = 0.31$ the measure of inefficiency σ increases for increasing λ. The efficiency is therefore the worst for the value $\lambda = \min(\alpha, 1-\alpha) = 0.31$. This

is in sharp contrast to the cases $\alpha = 0.5$ and $\alpha = 0.55$. In these two cases, there exists an intermediate value of $\lambda \in]0, \min(\alpha, 1-\alpha)[$ such that σ achieves its maximum. This is explained by the fact that, from a particular value of λ onward, both queues are less empty from time to time because λ is *high*. Hence, less slots are wasted. Finally, it is interesting to note that for the symmetric case $\alpha = 0.5$ the system is completely efficient ($\sigma = 0$) for $\lambda \to \alpha$.

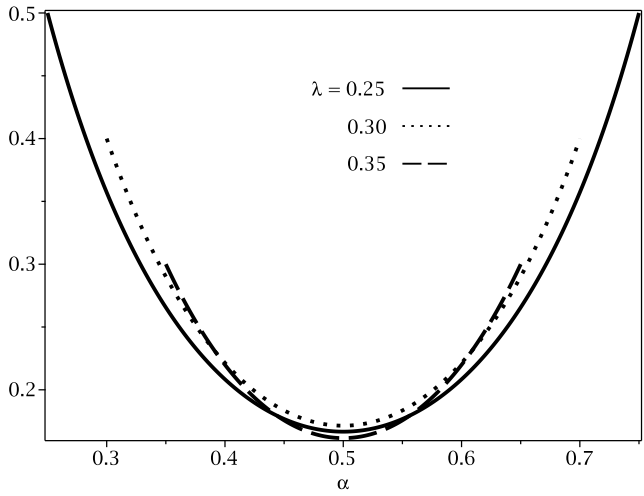


Figure 2.5: Measure of inefficiency σ versus α.

Note that since $\lambda_1 = \lambda_2 = \lambda$, the fractions of type-1 arrivals and type-2 arrivals in the overall traffic mix are equal. Hence, we expect that $\alpha = 0.5$ should be the optimal value for α in terms of efficiency. In Figure 2.5, we see that σ for varying α is symmetric around $\alpha = 0.5$ and that the minimum value is indeed found at $\alpha = 0.5$.

Next, we shift focus to the mean maximum system content. For this performance measure, we show that $\alpha = \frac{1}{2}$ is indeed always optimal. Consider $\mathsf{E}[\max(u_1, u_2)]$ as a function of α, with $\lambda < \alpha < 1-\lambda$ and $\lambda < 0.5$. This is a rational function with poles $1-\lambda$ and λ such that this is a well defined function for $\alpha \in]\lambda, 1-\lambda[$. It can easily be verified that, as a function of α, $\mathsf{E}[\max(u_1, u_2)]$ is convex and $\mathsf{E}[\max(u_1, u_2)] > 0$ for $\alpha \in]\lambda, 1-\lambda[$. Consequently, a unique minimizer in α exists. If we take the derivative of (2.103) with respect to α and require that this expression equals zero, we get

$$\frac{(1-\lambda)\lambda}{(1-\alpha-\lambda)^2} - \frac{(1-\lambda)\lambda}{(\alpha-\lambda)^2} = 0 \,.$$

Solving this equation to α yields

$$1 - \alpha - \lambda = \alpha - \lambda \,,$$

such that indeed $\alpha = \frac{1}{2}$. This proves that $\alpha^\star = \frac{1}{2}$ minimizes the mean maximum system content.

2.5 Geometric arrivals that are probabilistically routed to the two queues

Suppose that, during any slot, the *total* number of arrivals to the system is geometrically distributed with mean λ_T, i.e. the total number of arrivals during a slot is fully characterized according to the following PGF

$$A_T(z) = \frac{1}{1 + \lambda_T - \lambda_T z} \, .$$

Furthermore, an arriving customer is assumed to be of type-j with probability $\frac{\lambda_j}{\lambda_T}$, $j = 1, 2$ (with $\lambda_1 + \lambda_2 = \lambda_T$). The arriving customer is then routed to his designated queue. The joint PGF of the number of type-1 and type-2 arrivals can thus be written as $A_T\left(\frac{\lambda_1}{\lambda_T} z_1 + \frac{\lambda_2}{\lambda_T} z_2\right)$, which gives us

$$A(z_1, z_2) = \frac{1}{1 + \lambda_1 + \lambda_2 - \lambda_1 z_1 - \lambda_2 z_2} \, . \tag{2.104}$$

It is easy to verify that

$$\mathrm{cov}[a_1, a_2] = \lambda_1 \lambda_2 \, . \tag{2.105}$$

While the specific choice of arrival process in the previous section has led to independent queue contents, we obtain in this section that the system contents are statistically independent.

2.5.1 The marginal distributions $p_1(n)$ and $p_2(n)$

The type-1 and type-2 arrivals are geometrically distributed with mean λ_1 and λ_2 since

$$A(z, 1) = \frac{1}{1 + \lambda_1 - \lambda_1 z}$$

and

$$A(1, z) = \frac{1}{1 + \lambda_2 - \lambda_2 z} \, ,$$

respectively. From Section 1.5, we have that the PGF $U_1(z)$ describing the type-1 system content is given by

$$U_1(z) \triangleq U(z, 1) = \frac{\alpha - \lambda_1}{\alpha - \lambda_1 z} \, . \tag{2.106}$$

The reader will recognize the expression above as the PGF of a geometric distributed random variable with parameter $\frac{\lambda_1}{\alpha}$. Hence the pmf of u_1 is given as

$$\Pr[u_1 = n] = \left(1 - \frac{\lambda_1}{\alpha}\right)\left(\frac{\lambda_1}{\alpha}\right)^n , \quad n \geq 0 . \tag{2.107}$$

It easily follows that the radius of convergence τ_1 of $U_1(z)$ is equal to

$$\tau_1 = \frac{\alpha}{\lambda_1} . \tag{2.108}$$

For reasons of symmetry, we have that

$$U_2(z) \triangleq U(1, z) = \frac{1 - \alpha - \lambda_2}{1 - \alpha - \lambda_2 z} \tag{2.109}$$

and

$$\Pr[u_2 = n] = \left(1 - \frac{\lambda_2}{1 - \alpha}\right)\left(\frac{\lambda_2}{1 - \alpha}\right)^n , \quad n \geq 0 . \tag{2.110}$$

The radius of convergence τ_2 of $U_2(z)$ is given by

$$\frac{1 - \alpha}{\lambda_2} . \tag{2.111}$$

2.5.2 Areas of convergence

We have the same conclusions as in Section 2.3.2. The joint PGF $U(z_1, z_2)$ is analytic in the two polydiscs $|z_1| < \tau_1$, $|z_2| \leq 1$ and $|z_1| \leq 1$, $|z_2| < \tau_2$. Furthermore, the partial PGFs $U(z_1, 0)$ and $U(0, z_2)$ are analytic in $|z_1| < \tau_1$ and $|z_2| < \tau_2$, respectively.

2.5.3 Analysis of the kernel $K(z_1, z_2)$

In this subsection, we investigate the zeros of the kernel K. We have that

$$\begin{aligned}
&K(z_1, z_2) = 0 \\
\Leftrightarrow &\frac{-z_1^2 z_2 \lambda_1 - z_1 z_2^2 \lambda_2 + z_1 z_2 \lambda_1 + z_1 z_2 \lambda_2 + z_1 \alpha - z_2 \alpha + z_1 z_2 - z_1}{-\lambda_1 z_1 - \lambda_2 z_2 + \lambda_1 + \lambda_2 + 1} = 0 \\
\Leftrightarrow &- z_1^2 z_2 \lambda_1 - z_1 z_2^2 \lambda_2 + z_1 z_2 \lambda_1 + z_1 z_2 \lambda_2 + z_1 \alpha - z_2 \alpha + z_1 z_2 - z_1 = 0 \\
\Leftrightarrow &z_1 z_2 \left(-z_1 \lambda_1 - z_2 \lambda_2 + \lambda_1 + \lambda_2 + \frac{\alpha}{z_2} - \frac{\alpha}{z_1} + 1 - \frac{1}{z_2}\right) = 0 .
\end{aligned}$$

Letting

$$H_1(z_1) = \alpha + \lambda_1 - \lambda_1 z_1 - \frac{\alpha}{z_1} , \tag{2.112}$$

$$H_2(z_2) = 1 - \alpha + \lambda_2 - \lambda_2 z_2 - \frac{1-\alpha}{z_2} \,, \tag{2.113}$$

we see that

$$H_1(z_1) + H(z_2) = 0 \Rightarrow K(z_1, z_2) = 0 \,.$$

For complex values of z_1, we will only observe that for $z_1 = \sqrt{\frac{\alpha}{\lambda_1}} e^{\pm \mathrm{i}\theta}$, $H_1(z_1)$ is real and equal to

$$\alpha + \lambda_1 - \sqrt{\lambda_1 \alpha} 2 \cos(\theta) \,.$$

Moreover, we have that

$$\alpha + \lambda_1 - \sqrt{\lambda_1 \alpha} 2 \cos(\theta) > 0 \,, \tag{2.114}$$

since

$$\begin{aligned} \alpha + \lambda_1 - \sqrt{\lambda_1 \alpha} 2 \cos(\theta) &\geq \alpha + \lambda_1 - \sqrt{\lambda_1 \alpha} 2 \\ &= (\sqrt{\alpha} - \sqrt{\lambda_1})^2 \\ &> 0 \,. \end{aligned}$$

We are ready to prove the following lemma.

Lemma 2.6. *For values $z_1 = \sqrt{\frac{\alpha}{\lambda_1}} e^{\pm \mathrm{i}\theta}$, there is a unique $z_2 =: y(z_1) \in]0,1[$ such that*

$$H_1(z_1) + H_2(y(z_1)) = 0.$$

Proof. For $x \in [0,1]$, $H_2(x)$ increases monotonically from $-\infty$ at $x = 0$ to 0 at $x = 1$. Moreover, we have that

$$H_1 \left(\sqrt{\frac{\alpha}{\lambda_1}} e^{\pm \mathrm{i}\theta} \right) > 0 \,,$$

cf. (2.114). By virtue of the intermediate value theorem, there is a unique value z_2 in the interval $]0,1[$ such that

$$H_2(z_2) = -H_1 \left(\sqrt{\frac{\alpha}{\lambda_1}} e^{\pm \mathrm{i}\theta} \right) \,.$$

□

2.5.4 Analytic continuation of $U(z,0)$ and $U(0,z)$

The steps in order to determine $U(z,0)$ and $U(0,z)$ are analogous as in Section 2.4.4. First we notice that

$$y(z) = y(\bar{z}), \quad |z| = \sqrt{\frac{\alpha}{\lambda}},$$

for a proof we refer to Section 2.4.4. Moreover,

$$\sqrt{\frac{\alpha}{\lambda}} = \sqrt{\tau_1} < \tau_1$$

and $y(z) \in]0,1[$. Secondly, substitution of $(z, y(z))$ into (2.12) for values of z such that $|z|^2 = \frac{\alpha}{\lambda_1}$ gives us

$$(1-\alpha)(y(z)-1)zU(z,0) + \alpha(z-1)y(z)U(0,y(z)) = 0$$

and substitution of $(\bar{z}, y(z))$ for values of z such that $|z|^2 = \frac{\alpha}{\lambda_1}$ gives us

$$(1-\alpha)(y(z)-1)\bar{z}U(\bar{z},0) + \alpha(\bar{z}-1)y(z)U(0,y(z)) = 0.$$

Thirdly, eliminating $U(0, y(z))$ and rearranging terms yields

$$(\bar{z}-1)zU(z,0) = (z-1)\bar{z}U(\bar{z},0), \quad |z| = \sqrt{\frac{\alpha}{\lambda_1}}.$$

Finally, when multiplying both sides of the relation above by z and using the relations $z\bar{z} = \frac{\alpha}{\lambda_1} \Leftrightarrow |z| = |\bar{z}| = \sqrt{\frac{\alpha}{\lambda_1}}$ we find that

$$z\left(\frac{\alpha}{\lambda_1} - z\right)U(z,0) = \frac{\alpha}{\lambda_1}(z-1)U\left(\frac{\alpha}{\lambda_1}z^{-1}, 0\right), \quad |z| = \sqrt{\frac{\alpha}{\lambda_1}}. \tag{2.115}$$

Using equation (2.115), we now show that $U(z,0)$ admits a meromorphic continuation beyond $|z| < \tau_1$. We observe that the LHS of (2.115) is analytic in $|z| < \tau_1$. The RHS of (2.115) is analytic in $|z| > 1$, since $U(z,0)$ is analytic in $|z| < \tau_1$. Both sides of the equation have a common region of analyticity, namely the region $1 \leq |z| < \tau_1$. Hence, using analytic continuation we conclude that both sides of (2.115), in their respective regions of analyticity, are equal to the same unique entire function, say $h(z)$. The LHS of (2.115) is clearly bounded for $|z| < \tau_1$. Moreover, the RHS of (2.115) is bounded by a first degree polynomial for $|z| > 1$. Hence, the entire function $h(z)$ must be a first degree polynomial by Liouville's theorem. Hence, we have obtained that for $|z| < \tau_1$,

$$z\left(\frac{\alpha}{\lambda_1} - z\right)U(z,0) = h(z) = C_1 z + C_2$$

with C_j to be determined. Substituting $z = 0$ gives $C_2 = 0$. If we substitute $z_1 = 1$ and use that $U(1,0) = p_2(0) = 1 - \frac{\lambda_2}{1-\alpha}$, we obtain that

$$C_1 = \left(\frac{\alpha}{\lambda_1} - 1\right)\left(1 - \frac{\lambda_2}{1-\alpha}\right).$$

Summarized, we have obtained that $U(z,0)$ is given by the following expression

$$U(z,0) = \frac{(\alpha - \lambda_1)\left(1 - \frac{\lambda_2}{1-\alpha}\right)}{\alpha - \lambda_1 z}. \tag{2.116}$$

For reasons of symmetry, it can be shown that $U(0, z_2)$ is given by

$$U(0,z) = \frac{(1-\alpha-\lambda_2)\left(1-\frac{\lambda_1}{\alpha}\right)}{1-\alpha-\lambda_2 z}. \tag{2.117}$$

The reader might have noticed that $U(z_1, 0) = U_1(z_1)U_2(0)$ and that $U(0, z_2) = U_1(0)U_2(z_2)$. This suggests that u_1 and u_2 are statistically independent.

2.5.5 The joint distribution $p(n, m)$

If we substitute the expressions (2.116) and (2.117) for $U(z, 0)$ and $U(0, z)$, respectively, into the functional equation for $U(z_1, z_2)$, we get

$$U(z_1,z_2) = \frac{(\alpha-\lambda_1)(1-\alpha-\lambda_2)}{(\alpha-\lambda_1 z_1)(1-\alpha-\lambda_2 z_2)} \, . \tag{2.118}$$

We can conclude that $U(z_1, z_2) = U(z_1, 1)U(1, z_2)$, i.e., the system contents are statistically independent. We emphasize that this is a striking result. However, we cannot intuitively explain this result.

The joint pmf $p(n, m)$ of type-1 and type-2 system contents is easily obtained because the joint pmf can be factorized as

$$p(n,m) = p_1(n)p_2(m) \, .$$

We obtain that

$$p(n,m) = \frac{(\alpha-\lambda_1)(1-\alpha-\lambda_2)}{\alpha(1-\alpha)} \left(\frac{\lambda_1}{\alpha}\right)^n \left(\frac{\lambda_2}{1-\alpha}\right)^m \, . \tag{2.119}$$

Generalization: other PGFs $A(z_1, z_2)$ such that the system contents are statistically independent

We show that the solution of the functional equation (2.12) is given by

$$U(z_1,z_2) = U_1(z_1)U_2(z_2) \, , \tag{2.120}$$

as soon as the joint arrival PGF $A(z_1, z_2)$ satisfies

$$A(z_1,z_2) = \frac{A_1(z_1)A_2(z_2)}{A_1(z_1)+A_2(z_2)-A_1(z_1)A_2(z_2)} \, . \tag{2.121}$$

Indeed, consider an arrival PGF $A(z_1, z_2)$ for which (2.121) holds. We check if $U(z_1, z_2) = U_1(z_1)U_2(z_2)$ satisfies the functional equation (2.12). In a first step, we notice that (2.120) implies

$$U(z_1,z_2) = \frac{(\alpha-\lambda_1)(1-\alpha-\lambda_2)(z_1-1)(z_2-1)A_1(z_1)A_2(z_2)}{(z_1-A_1(z_1)(\alpha+(1-\alpha)z_1))(z_2-A_2(z_2)(1-\alpha+(1-\alpha)z_2))} \, , \tag{2.122}$$

whence,

$$U(z_1,0) = \frac{(\alpha-\lambda_1)(z_1-1)A_1(z_1)}{z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1)} \frac{1-\alpha-\lambda_2}{1-\alpha} , \tag{2.123}$$

$$U(0,z_2) = \frac{(1-\alpha-\lambda_2)(z_2-1)A_2(z_2)}{z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2)} \frac{\alpha-\lambda_1}{\alpha} . \tag{2.124}$$

Substituting these two expressions for $U(z_1,0)$ and $U(0,z_2)$ into the RHS of equation (2.12) yields

$$\begin{aligned}
&A(z_1,z_2)((1-\alpha)(z_2-1)z_1U(z_1,0) + \alpha(z_1-1)z_2U(0,z_2)) \\
&= A(z_1,z_2)\left(\frac{(z_2-1)z_1(\alpha-\lambda_1)(z_1-1)A_1(z_1)(1-\alpha-\lambda_2)}{z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1)}\right. \\
&\quad \left. + \frac{(z_1-1)z_2(1-\alpha-\lambda_2)(z_2-1)A_2(z_2)(\alpha-\lambda_1)}{z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2)}\right) \\
&= A(z_1,z_2)(z_1-1)(z_2-1)(\alpha-\lambda_1)(1-\alpha-\lambda_2) \\
&\quad \times \left(\frac{z_1A_1(z_1)}{z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1)} + \frac{z_2A_2(z_2)}{z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2)}\right)
\end{aligned} .$$

Now notice that

$$\begin{aligned}
&\frac{z_1A_1(z_1)}{z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1)} + \frac{z_2A_2(z_2)}{z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2)} \\
&= \frac{z_1z_2(A_1(z_1)+A_2(z_2)-A_1(z_1)A_2(z_2)) - A_1(z_1)A_2(z_2)((1-\alpha)z_1+\alpha z_2)}{(z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1))(z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2))} \\
&= \frac{z_1z_2\frac{A_1(z_1)A_2(z_2)}{A(z_1,z_2)} - A_1(z_1)A_2(z_2)((1-\alpha)z_1+\alpha z_2)}{(z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1))(z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2))} ,
\end{aligned}$$

where in the last equality we used (2.121). Hence

$$\begin{aligned}
&A(z_1,z_2)((1-\alpha)(z_2-1)z_1U(z_1,0) + \alpha(z_1-1)z_2U(0,z_2)) \\
&= \frac{A(z_1,z_2)(z_1-1)(z_2-1)(\alpha-\lambda_1)(1-\alpha-\lambda_2)}{(z_1 - A_1(z_1)(\alpha+(1-\alpha)z_1))(z_2 - A_2(z_2)(1-\alpha+(1-\alpha)z_2))} \\
&\quad \times \left(z_1z_2\frac{A_1(z_1)A_2(z_2)}{A(z_1,z_2)} - A_1(z_1)A_2(z_2)((1-\alpha)z_1+\alpha z_2)\right) \\
&= \frac{A(z_1,z_2)}{A_1(z_1)A_2(z_2)}U(z_1,z_2) \\
&\quad \times \left(z_1z_2\frac{A_1(z_1)A_2(z_2)}{A(z_1,z_2)} - A_1(z_1)A_2(z_2)((1-\alpha)z_1+\alpha z_2)\right) \\
&= U(z_1,z_2)\left(z_1z_2 - A(z_1,z_2)((1-\alpha)z_1+\alpha z_2)\right) ,
\end{aligned}$$

such that indeed (2.122) is a solution of equation (2.12). Moreover, (2.122) is a joint PGF since it is the product of two proper marginal PGFs and hence our claim is proved.

We will now introduce a large class of PGFs $A(z_1, z_2)$ that satisfy condition (2.121). Consider

$$A(z_1, z_2) = \frac{1}{1 + \mu_1 + \mu_2 - \mu_1 L_1(z_1) - \mu_2 L_2(z_2)} , \tag{2.125}$$

where $L_1(z)$ and $L_2(z)$ are PGFs. Note that the case where

$$L_1(z) = L_2(z) = z$$

corresponds to the arrival PGF (2.104). Now

$$A_1(z) = \frac{1}{1 + \mu_1 - \mu_1 L_1(z)}$$

and

$$A_2(z) = \frac{1}{1 + \mu_2 - \mu_2 L_2(z)} .$$

Hence,

$$A_1(z_1) + A_2(z_2) = \frac{2 + \mu_1 + \mu_2 - \mu_1 L_1(z_1) - \mu_2 L_2(z_2)}{(1 + \mu_1 - \mu_1 L_1(z_1))(1 + \mu_2 - \mu_2 L_2(z_2))} ,$$

such that

$$\begin{aligned} A_1(z_1) + A_2(z_2) - A(z_1)A(z_2) &= \frac{1 + \mu_1 + \mu_2 - \mu_1 L_1(z_1) - \mu_2 L_2(z_2)}{(1 + \mu_1 - \mu_1 L_1(z_1))(1 + \mu_2 - \mu_2 L_2(z_2))} \\ &= \frac{A_1(z_1)A_2(z_2)}{A(z_1, z_2)} . \end{aligned}$$

In conclusion, PGFs of the form (2.125) will lead to a product-form solution for the joint PGF $U(z_1, z_2)$ of the system contents.

2.5.6 The marginal distribution $p_T(n)$

The marginal distribution $p_T(n)$ of the total system content can be easily obtained. Because the system contents are statistically independent, it holds that

$$p_T(n) = \sum_{k=0}^{n} p_1(k) p_2(n-k) .$$

Whence,

$$\begin{aligned} p_T(n) = &\frac{(\alpha - \lambda_1)(1 - \alpha - \lambda_2)\lambda_1}{\alpha((1-\alpha)\lambda_1 - \alpha\lambda_2)} \left(\frac{\lambda_1}{\alpha}\right)^n \\ &- \frac{(\alpha - \lambda_1)(1 - \alpha - \lambda_2)\lambda_2}{(1-\alpha)((1-\alpha)\lambda_1 - \alpha\lambda_2)} \left(\frac{\lambda_2}{1-\alpha}\right)^n . \end{aligned} \tag{2.126}$$

2.5.7 Calculation of numerical characteristics

Moments

Since the type-1 and type-2 system contents are geometrically distributed, the marginal moments of the type-1 and type-2 system contents can readily be obtained. It follows that

$$\mathsf{E}[u_1] = \frac{\lambda_1}{\alpha - \lambda_1} \tag{2.127}$$

$$\mathrm{var}[u_1] = \frac{\alpha\lambda_1}{(\alpha - \lambda_1)^2} \ . \tag{2.128}$$

Likewise, we have that

$$\mathsf{E}[u_2] = \frac{\lambda_2}{1 - \alpha - \lambda_2} \tag{2.129}$$

$$\mathrm{var}[u_2] = \frac{(1-\alpha)\lambda_2}{(1 - \alpha - \lambda_2)^2} \ . \tag{2.130}$$

Finally, since the system contents are statistically independent, it follows that

$$\mathrm{cov}[u_1, u_2] = 0 \ , \tag{2.131}$$

and

$$\begin{aligned} \mathrm{var}[u_T] &= \mathrm{var}[u_1] + \mathrm{var}[u_2] \\ &= \frac{\alpha\lambda_1}{(\alpha - \lambda_1)^2} + \frac{(1-\alpha)\lambda_2}{(1 - \alpha - \lambda_2)^2} \ . \end{aligned} \tag{2.132}$$

A measure of inefficiency

As in Section 2.3 and Section 2.4, the probability that the server selects an empty queue while the non-selected queue is non-empty is denoted by σ and can be calculated as

$$\sigma = p_2(0)(1-\alpha) + p_1(0)\alpha - p(0,0) \ .$$

If we substitute the expressions for $p_1(0)$, $p_2(0)$ and $p(0,0)$, we find that

$$\sigma = \frac{(1-\alpha)^2\lambda_1 + \alpha^2\lambda_2 - \lambda_1\lambda_2}{\alpha(1-\alpha)} \ . \tag{2.133}$$

The mean maximum system content

The mean maximum system content is in this case easily computed since

$$\lim_{k\to\infty} \Pr[\max(u_{1,k}, u_{2,k}) > L] = 1 - \lim_{k\to\infty} \Pr[\max(u_{1,k}, u_{2,k}) \leq L]$$

$$
\begin{aligned}
&= 1 - \lim_{k\to\infty} \Pr[u_{1,k} \le L] \lim_{k\to\infty} \Pr[u_{2,k} \le L] \\
&= 1 - \left(1 - \left(\frac{\lambda_1}{\alpha}\right)^{L+1}\right)\left(1 - \left(\frac{\lambda_2}{1-\alpha}\right)^{L+1}\right) \\
&= \left(\frac{\lambda_1}{\alpha}\right)^{L+1} + \left(\frac{\lambda_2}{1-\alpha}\right)^{L+1} - \left(\frac{\lambda_1\lambda_2}{\alpha(1-\alpha)}\right)^{L+1} .
\end{aligned}
$$

Whence,

$$
\begin{aligned}
\mathsf{E}[\max(u_1, u_2)] &= \sum_{L=0}^{\infty} \lim_{k\to\infty} \Pr[\max(u_{1,k}, u_{2,k}) > L] \\
&= \frac{\lambda_1}{\alpha - \lambda_1} + \frac{\lambda_2}{1-\alpha-\lambda_2} - \frac{\lambda_1\lambda_2}{\alpha(1-\alpha) - \lambda_1\lambda_2} . \qquad (2.134)
\end{aligned}
$$

We have studied the influence of α, λ_1 and λ_2 on the mean maximum system content and on the measure σ. However, the behavior is very similar of that of Section 2.3. Therefore the numerical results are omitted.

2.6 Concluding remarks

In this chapter, we have analyzed a discrete-time two-class queueing model with randomly alternating service, single-slot service times and infinite waiting room. Various assumptions for the arrival distribution have been made in order to present an exact analysis. We will discus asymptotic and approximation techniques to analyze this model in the next chapters. The usefulness of this chapter is twofold.

Firstly, despite the fact that the results obtained in this chapter are only valid for specific arrival distributions, these are useful in their own right. This is because the obtained expressions make it possible to demonstrate (in a relatively simple way) the impact of the scheduling discipline on the performance of the queueing model.

Secondly, we demonstrated the difficulty in solving functional equations like (2.1) for relatively simple kernels $K(z_1, z_2)$. The insight gained in the used method will help us in the following chapter. The asymptotic analysis used in the next chapter will be highly based on the analysis of Section 2.3.

3

Asymptotic analysis: tail asymptotics

In the previous chapter, we have analyzed our queueing model with randomly alternating service discipline under various assumptions for the arrival distribution. The closed-form analysis of the general queueing model, i.e. where the numbers of arrivals per time slot have a general joint distribution, proves to be unfeasible. Therefore, we shift focus to the asymptotic behavior of the joint stationary distribution $p(i,j)$ of the system contents (under less severe restrictions on the arrival distribution). To make this chapter self-contained, we briefly repeat the exact definitions of our queueing model as described in Section 1.4. We consider a discrete-time queueing model with two infinite-sized queues and one server. There are two types of customers, where the type of the customer corresponds to a specific queue. Each customer has a service time of a single slot. In each time slot, the server is available to queue-1 with probability α and to queue-2 with probability $1-\alpha$. When an empty queue is chosen, no service occurs in that slot, even when the other queue is non-empty. The number of type-j arrivals in slot k is denoted by $a_{j,k}$, $j=1,2$. The joint PGF of $a_{1,k}$ and $a_{2,k}$ is denoted by $A(z_1,z_2)$. Finally, we denoted the type-j system content at the beginning of slot k by $u_{j,k}$.

To better describe what we are trying to achieve in this chapter, consider the special case of the arrival distribution in Section 2.3. The joint pmf $p(i,j)$ of the system contents in this case was given by, cf. (2.62)

$$p(i,j) = c_1\tau_1^{-i}\tau_T^{-j} + c_2\tau_T^{-i}\tau_2^{-j} + c_3\tau_T^{-(i+j)} \quad (i,j \geq 1)\,, \tag{3.1}$$

with c_1, c_2 and c_3 constants that are independent of i and j. Recall that $\tau_T = \tau_1\tau_2$, cf. (2.23). From (3.1), it is clear that, if we are only interested in $p(i,j)$ for large i and fixed j, we can only keep the term with τ_1^{-i} and neglect the other terms. We then have

$$p(i,j) \sim c_1\tau_T^{-j}\tau_1^{-i}, \quad \text{as } i \to \infty\,, \tag{3.2}$$

where we write $f(i) \sim g(i)$ if $\lim_{i\to\infty} f(i)/g(i) = 1$.

More formally, the tail asymptotic problem for a discrete non-negative random variable X is that of finding an approximation for the probability $\Pr[X = n]$ by a much simpler function of n, say $h(n)$, such that

$$\Pr[X = n] \sim h(n), \quad \text{as } n \to \infty .$$

Such a function h dictates the decay of $\Pr[X = n]$, for large n.

For a random vector (X, Y), the formulation of the tail asymptotic problem may differ since a joint distribution has several directions. In this chapter, we will focus on the asymptotics along a coordinate direction. That is, for every fixed m we search for a simple function h_m such that

$$\Pr[X = n, Y = m] \sim h_m(n), \quad \text{as } n \to \infty .$$

The results in this chapter are closely connected to the asymptotic analysis of two-dimensional Markov chains. The main difficulty in this analysis is that, in general, no analytical expression for the joint PGF is available. Despite this difficulty, the tail asymptotic problem for two-dimensional queueing models has been investigated extensively for nearest-neighbor random walks the last two decades, see for example [82–87] and references therein. In [88], Kobayashi and Miyazawa obtain tail asymptotic results for the marginal stationary distributions of a nearest-neighbor random walk in $\{0, 1, \ldots\}^2$. They have established that these distributions exhibit the following law: $\sim cn^{\nu}\tau^{-n}$, with $\tau > 1$, $\nu \in \{-\frac{3}{2}, -\frac{1}{2}, 0, 1\}$ and c unspecified. Their approach is comparable to the method employed in [86, 89]. In essence, both methods succeed in determining the location and the nature of the singularities of the boundary functions $\Phi(z_1, 0)$ and $\Phi(0, z_2)$ via the functional equation (2.1). Once these singularities are obtained, the singularity analysis of the marginal PGFs $\Phi(z_1, 1)$, $\Phi(1, z_2)$ and $\Phi(z, z)$ readily follows. Finally, the tail asymptotics follow from singularity analysis.

It is worth noting that the tail asymptotic problem can be limited to that of finding only the decay rate β, defined by

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr[X = n] = -\beta$$

provided that this limit exists. Note that if $\Pr[X = n] \sim cn^{\nu}\tau^{-n}$, then it follows that the tail decay rate β is given by $\ln(\tau)$.

In [85], the tail decay rate is obtained for the marginal distributions of random walks in the quarter plane for which the one-step displacements are not restricted to neighboring states. In [90], the tail decay rates for a nearest-neighbor random walk in the quarter plane that is modulated by a background process are investigated. This modulation means that the transitions of the random walk depend on the state of a background process. Assuming stability

conditions for the background process, Ozawa and Kobayashi determine the dominant singularity of the boundary functions $\Phi(z_1, 0)$ and $\Phi(0, z_2)$ in [91]. The general study of exponential decay in rare events is that of large-deviations theory. We refer to [92, 93] for good introductions to this theory.

For clarity, we explicitly state the objectives of this chapter:

1. We investigate under which condition on the arrival process, the joint distribution $p(i, j)$ (for fixed i or j) has an asymptotic geometric behavior.
2. We compute the tail asymptotics under this condition.

For convenience, let the radius of convergence of the infinite series of the PGF of the number of arrivals of type-j customers in a slot

$$A_j(z) = \sum_{n=0}^{\infty} a_j(n) z^n \ ,$$

be denoted by $\mathcal{R}_j$ $(j = 1, 2)$. We consider well-behaved functions $A_j(z)$, therefore we assume throughout this chapter that

Assumption 3.1. $\mathcal{R}_j > 1 \quad$ *and* $\quad \lim_{z \to \mathcal{R}_j} A_j(z) = +\infty \ , \ j = 1, 2 \ .$

This class of discrete probability distributions contains many well-known distributions, for example: the binomial distribution, the geometric distribution and the Poisson distribution satisfy Assumption 3.1. In fact, distributions whose PGF is entire ($\mathcal{R}_j = +\infty$) or whose PGF is a rational function are the most common examples of arrival distributions for which Assumption 3.1 holds.

The method we apply to obtain asymptotic formulas, is singularity analysis of the (partial) PGFs describing the system contents. There is general correspondence between the asymptotic expansion of a generating function near its dominant singularities and the asymptotic expansion of the coefficients of the generating functions. For a detailed explanation of this method we refer the reader to [11, Part B] or [94, Sect. 3]. The generating functions that will be subjected to singularity analysis in this chapter, have poles as dominant singularities (as we will show). Therefore, the following theorem will be sufficient.

Theorem 3.1. *Let $X(z) = \sum_{i=0}^{\infty} x(i) z^i$ be a meromorphic function at all points of the closed disc $|z| \leq R$ with poles at points $\beta_1, \beta_2, \ldots, \beta_n$ with multiplicity $m_1, m_2, \ldots, m_n$ respectively. Assume that $X(z)$ is analytic at all points of $|z| = R$ and at $z = 0$. In a punctured disk around β_l, $X(z)$ has the Laurent expansion*

$$X(z) = \sum_{k=0}^{\infty} d_{l,k} (z - \beta_l)^k + \sum_{k=1}^{m_l} \frac{b_{l,k}}{(z - \beta_l)^k} \ ,$$

with $d_{l,k}$ $(k = 0, 1, \ldots)$ and $b_{l,k}$ $(k = 1, \ldots, m_l)$ complex numbers. Then,

$$x(i) \sim \sum_{l=1}^{n} \sum_{k=1}^{m_l} \binom{i + k - 1}{i} (-1)^k b_{l,k} \beta_l^{-(i+k)} \ . \tag{3.3}$$

Proof. See for example the proof of Theorem 5.2.1 in [95]. □

In the special case that there is only one pole and the multiplicity of this pole is 1 (we call this a simple pole), then (3.3) simplifies to

$$x(i) \sim -b_{1,1}\beta_1^{-(i+1)} ,$$

with $b_{1,1}$ the residue of $X(z)$ at β_1, given by

$$\operatorname*{res}_{z=\beta_1} X(z) = \lim_{z\to\beta_1} (z-\beta_1)X(z) .$$

This particular case is in fact the only case we will encounter throughout this chapter. Without going into great detail, we want to remark that Theorem 3.1 is the simplest application of singularity analysis, since the dominant singularities are in this case poles. However, singularity analysis is also applicable to functions whose singular expansion involves fractional powers and logarithms. For further details of this theory, we refer to the classic book [11].

The remainder of this Chapter is organized as follows. In Section 3.1 we repeat the most important definitions and give several preliminary results. In Section 3.2 we introduce the major conditions on the PGF $A(z_1, z_2)$ such that a geometric tail behavior is obtained. This is followed by the derivation of these conditions in Section 3.3. The tail asymptotics of the complete joint distribution $p(i, j)$ are provided in Section 3.4. Finally, in Section 3.5 we consider the special case of independent arrivals in the queues, which allows for a supplementary analysis.

3.1 Preliminaries

For definiteness, we repeat some of the most important definitions of Chapter 2. Under the assumption that the system can reach a steady state, we defined

$$p(i,j) = \lim_{k\to\infty} \mathsf{P}[u_{1,k} = i, u_{2,k} = j], \quad i,j \geq 0 \tag{3.4}$$

$$p_1(i) = \sum_{j=0}^{\infty} p(i,j), \quad i \geq 0 \tag{3.5}$$

$$p_2(j) = \sum_{i=0}^{\infty} p(i,j), \quad j \geq 0 . \tag{3.6}$$

We also defined the following PGFs

$$U(z_1,z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p(i,j) z_1^i z_2^j \tag{3.7}$$

$$U_1(z) = \sum_{i=0}^{\infty} p_1(i) z^i \tag{3.8}$$

$$U_2(z) = \sum_{j=0}^{\infty} p_2(j) z^j \ , \tag{3.9}$$

which are the joint PGF of the numbers of type-1 and type-2 customers in the system, the marginal PGF of the number of type-1 and the marginal PGF of the number of type-2 customers in the system respectively. Of course, we have that $U_1(z) = U(z,1)$ and $U_2(z) = U(1,z)$. The joint PGF satisfies the following functional equation, cf. (2.12),

$$\begin{aligned} K(z_1,z_2)U(z_1,z_2) &= A(z_1,z_2) \\ &\quad \times [(1-\alpha)(z_2-1)z_1 U(z_1,0) + \alpha(z_1-1)z_2 U(0,z_2)] \ , \end{aligned} \tag{3.10}$$

with

$$K(z_1,z_2) = z_1 z_2 - [(1-\alpha)z_1 + \alpha z_2] A(z_1,z_2) \ . \tag{3.11}$$

We now define the following new partial probability generating functions:

$$P_{1,n}(z) = \sum_{i=0}^{\infty} p(i,n) z^i, \quad n = 0,1,\dots \ , \tag{3.12}$$

$$P_{2,n}(z) = \sum_{j=0}^{\infty} p(n,j) z^j, \quad n = 0,1,\dots \tag{3.13}$$

as the partial PGF of the number of type-1 customers in the system when there are n type-2 customers in the system and the partial PGF of the number of type-2 customers in the system when there are n type-1 customers present, respectively. Notice that $P_{1,0}(z) = U(z,0)$ and $P_{2,0}(z) = U(0,z)$. Moreover, we can expand $U(z_1,z_2)$ as

$$U(z_1,z_2) = \sum_{n=0}^{\infty} P_{1,n}(z_1) z_2^n \ , \tag{3.14}$$

or

$$U(z_1,z_2) = \sum_{n=0}^{\infty} P_{2,n}(z_2) z_1^n \ . \tag{3.15}$$

3.1.1 Recurrence relations

In this subsection, we show that the functions $P_{1,n}(z)$ and $P_{2,n}(z)$ satisfy a particular recurrence relation. This can be established by writing down the balance equations and then taking the corresponding z-transform. However, we follow a different approach. We substitute the expansions (3.14) (or (3.15)) into (3.10) and equate coefficients in z_2 (or in z_1).

Let us define

$$A_{1,n}(z) \triangleq \sum_{i=0}^{\infty} a(i,n) z^i \tag{3.16}$$

as the partial PGF of the number of type-1 arrivals in a slot with n type-2 customers arrivals. Notice that

$$A(z_1, z_2) = \sum_{n=0}^{\infty} A_{1,n}(z_1) z_2^n \,. \tag{3.17}$$

We now first establish a recurrence relation for $P_{1,n}(z)$. If we substitute (3.14) and (3.17) into (3.10), we get

$$\begin{aligned}
&\sum_{n=0}^{\infty} z_1 P_{1,n}(z_1) z_2^{n+1} - \sum_{n=0}^{\infty} \left((1-\alpha) z_1 \sum_{k=0}^{n} A_{1,n-k}(z_1) P_{1,k}(z_1) \right) z_2^n \\
&- \sum_{n=0}^{\infty} \left(\alpha \sum_{k=0}^{n} A_{1,n-k}(z_1) P_{1,k}(z_1) \right) z_2^{n+1} = \sum_{n=0}^{\infty} A_{1,n}(z_1)(1-\alpha) z_1 P_{1,0}(z_1) z_2^{n+1} \\
&- \sum_{n=0}^{\infty} A_{1,n}(z_1)(1-\alpha) z_1 P_{1,0}(z_1) z_2^n + \sum_{n=0}^{\infty} \alpha (z_1 - 1) \sum_{k=0}^{n} A_{1,n-k}(z_1) P_{1,k}(0) z_2^{n+1} \,.
\end{aligned} \tag{3.18}$$

Hence, by equating coefficients in z_2 we obtain that for $n = 0, 1, \ldots$:

$$\begin{aligned}
z P_{1,n}(z) = {} & \alpha \sum_{j=0}^{n} A_{1,n-j}(z) [P_{1,j}(z) + (z-1) P_{1,j}(0)] \\
& + (1-\alpha) z A_{1,n}(z) P_{1,0}(z) + (1-\alpha) z \sum_{j=0}^{n} A_{1,n-j}(z) P_{1,j+1}(z),
\end{aligned} \tag{3.19}$$

For reasons of symmetry, we have the following recurrence relation for $P_{2,n}(z)$:

$$\begin{aligned}
z P_{2,n}(z) = {} & \alpha z A_{2,n}(z) P_{2,0}(z) + \alpha z \sum_{i=0}^{n} A_{2,n-i}(z) P_{2,i+1}(z) \\
& + (1-\alpha) \sum_{i=0}^{n} A_{2,n-i}(z) [P_{2,i}(z) + (z-1) P_{2,i}(0)] \,,
\end{aligned} \tag{3.20}$$

where we defined

$$A_{2,n}(z) \triangleq \sum_{j=0}^{\infty} a(n,j) z^j \tag{3.21}$$

as the partial PGF of the number of type-2 arrivals in a slot with n type-1 customers arrivals.

If we can obtain the dominant singularities of $P_{1,0}(z)$ and $P_{2,0}(z)$, we see from equations (3.19) and (3.20) that we can also recursively obtain the dominant singularities of $P_{1,n}(z)$ and $P_{2,n}(z)$, $n = 1, 2, \ldots$. In Chapter 2, we considered three special cases for $A(z_1, z_2)$ and for each of these special cases it

was possible to determine the dominant singularity of $U(z,0) = P_{1,0}(z)$ and $U(0,z) = P_{2,0}(z)$ without much difficulty. In fact, it turned out that the dominant singularities of $U(z,0)$ and $U(0,z)$ coincide with the dominant singularities of $U_1(z)$ and $U_2(z)$, respectively. Therefore, we first investigate the dominant singularities of $U_1(z)$ and $U_2(z)$.

3.1.2 Asymptotic analysis of $U_1(z)$ and $U_2(z)$

A typical feature of the queueing model as described in Section 1.4 is the fact that the expression for the marginal PGFs of the number of type-1 and type-2 customers in the system is known, see also Section 1.5. This information can be used to obtain a lower bound for the radius of convergence of $U(z_1,0)$ and $U(0,z_2)$, as we did in Section 2.3.2 for a specific arrival process. For the sake of completeness and since this is an essential element of this chapter, we repeat the argument of Section 1.5. The reasoning is as follows. We investigate for which values of z the infinite series

$$U(z,0) = \sum_{i=0}^{\infty} p(i,0) z^i$$

converges. We observe that for every $i \in \mathbb{N}$

$$\begin{aligned} p(i,0) &\leq p(i,0) + p(i,1) + p(i,2) + \dots \\ &= \sum_{j=0}^{\infty} p(i,j) \\ &= p_1(i) \ . \end{aligned}$$

Hence, the radius of convergence of $U_1(z)$ is a lower bound for the radius of convergence of $U(z,0)$. Analogously, the radius of convergence of $U_2(z)$ is a lower bound for the radius of convergence of $U(0,z)$. For this reason, we feel it is useful to first determine the radius of convergence of the marginal PGFs $U_1(z_1)$ and $U_2(z_2)$.

The formulas for $U_1(z_1)$ and $U_2(z_2)$ are given by

$$U_1(z_1) = U(z_1,1) = \frac{(z_1-1)A_1(z_1)(\alpha-\lambda_1)}{K(z_1,1)} \ , \tag{3.22}$$

$$U_2(z_2) = U(1,z_2) = \frac{(z_2-1)A_2(z_2)(1-\alpha-\lambda_2)}{K(1,z_2)} \ . \tag{3.23}$$

Under Assumption 3.1, we have the following well-known result [40].

Lemma 3.1.

1. *Equation $K(z_1,1) = 0$ has exactly two real positive roots inside the interval $[0, \mathcal{R}_1[$, 1 and say τ_1, such that $1 < \tau_1 < \mathcal{R}_1$.*

2. $K^{(1)}(\tau_1, 1) < 0$.

3. $K(z_1, 1)$ *has no other zeros with the same absolute value as* τ_1.

Although this lemma is well-known (and thus not new), we will give a proof because we will use the same methodology again in Section 3.5.1.

Proof of Lemma 3.1. (1) and (2): Recall that $A_1(z_1)$ is a PGF. Consequently, for $z_1 \in [0, \mathcal{R}_1[$, $A_1(z_1)$ is a power series with non-negative coefficients. Hence,

$$\begin{aligned} K^{(11)}(z_1, 1) &= -2(1-\alpha)A_1'(z_1) - ((1-\alpha)z_1 + \alpha)A_1''(z_1) \\ &< 0 \,, \end{aligned}$$

for $z_1 \in [0, \mathcal{R}_1[$. Furthermore, it is easy to see that $K(1, 1) = 0$ and that

$$K^{(1)}(1, 1) = \alpha - \lambda_1 > 0 \,.$$

The latter follows from the stability condition. Furthermore, by Assumption 3.1, we have that

$$K(z_1, 1) \to -\infty \quad \text{as } z_1 \to \mathcal{R}_1 \,.$$

We can conclude that $K(z_1, 1)$ has a unique zero in $]1, \mathcal{R}_1[$. This proves part 1 and 2 of the lemma.

(3): Consider z_1 for which $|z_1| = \tau_1$, $z_1 \neq \tau_1$. For these values of z_1 we have that

$$|(1-\alpha)z_1 + \alpha| < (1-\alpha)\tau_1 + \alpha, \,,$$

it follows that for $|z_1| = \tau_1$, $z_1 \neq \tau_1$,

$$\begin{aligned} |(1-\alpha)z_1 + \alpha)A_1(z_1)| &< ((1-\alpha)|z_1| + \alpha)A_1(|z_1|) \\ &= ((1-\alpha)\tau_1 + \alpha)A_1(\tau_1) \\ &= \tau_1 \\ &= |z_1| \,. \end{aligned}$$

Hence 3 is proven. □

In other words, the graph of the function $K(x, 1)$ resembles that of a parabola that opens downward and which intersects the x axis at $x = 1$ and $x = \tau_1$ (with $\tau_1 > 1$), see Figure 3.1.

Note that the value τ_1 is thus the unique zero of $K(z, 1)$ in $]1, \mathcal{R}_1[$, i.e.

$$\tau_1 = ((1-\alpha)\tau_1 + \alpha)A_1(\tau_1), \quad 1 < \tau_1 < \mathcal{R}_1 \,. \tag{3.24}$$

We now have the following theorem.

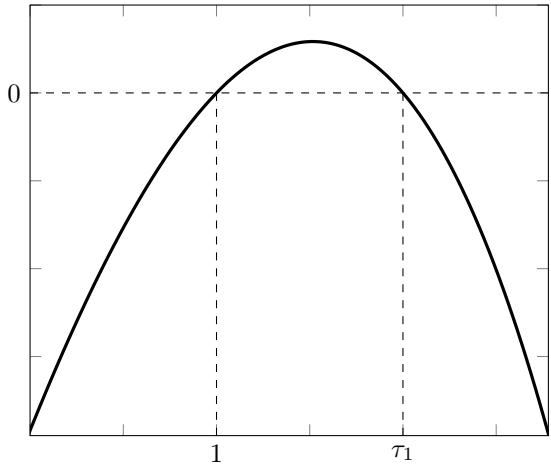


Figure 3.1: Illustration of the function $K(x, 1)$ for positive real x.

Theorem 3.2. *The asymptotics of the distribution of the number of type-1 customers in the system are given by*

$$p_1(n) \sim C_1 \tau_1^{-(n+1)} , \tag{3.25}$$

with

$$C_1 = \frac{(\tau_1 - 1)A_1(\tau_1)(\alpha - \lambda_1)}{(1-\alpha)A_1(\tau_1) + [(1-\alpha)\tau_1 + \alpha]A_1'(\tau_1) - 1} . \tag{3.26}$$

Proof. From Lemma 3.1, we know that τ_1 is the unique dominant singularity of $U_1(z)$ and that this is necessarily a simple pole. Furthermore, we have that

$$\begin{aligned}
\lim_{z\to\tau_1} (z - \tau_1)U_1(z) &= \lim_{z\to\tau_1} \frac{(z - \tau_1)(z_1 - 1)A_1(z_1)(\alpha - \lambda_1)}{K(z_1, 1)} \\
&= \frac{(\tau_1 - 1)A_1(\tau_1)(\alpha - \lambda_1)}{K^{(1)}(\tau_1, 1)} \\
&= -\frac{(\tau_1 - 1)A_1(\tau_1)(\alpha - \lambda_1)}{(1-\alpha)A_1(\tau_1) + [(1-\alpha)\tau_1 + \alpha]A_1'(\tau_1) - 1} .
\end{aligned}$$

By virtue of Theorem 3.1, the theorem is now proven. □

We now turn our attention to the type-2 customers. We have the following symmetrical result, compared to Lemma 3.1.

Lemma 3.2.

1. *Equation $K(1, z_2) = 0$ has exactly two real positive roots, 1 and say τ_2, such that $1 < \tau_2 < \mathcal{R}_2$.*

2. $K^{(2)}(1, \tau_2) < 0$.

3. *$K(1, z_2)$ has no other zeros with the same absolute value as τ_2.*

Proof. Analogously as in the proof of Lemma 3.1. □

The value τ_2 is thus the unique zero of $K(1, z)$ in $]1, \mathcal{R}_2[$, i.e.

$$\tau_2 = (1 - \alpha + \alpha\tau_2)A_2(\tau_2), \quad 1 < \tau_2 < \mathcal{R}_1 \, . \tag{3.27}$$

Finally, we also have the following theorem.

Theorem 3.3. *The asymptotics of the distribution of the number of type-2 customers in the system are given by*

$$p_2(n) \sim C_2 \tau_2^{-(n+1)} \, , \tag{3.28}$$

with

$$C_2 = \frac{(\tau_2 - 1)A_2(\tau_2)(1 - \alpha - \lambda_2)}{\alpha A_2(\tau_2) + [1 - \alpha + \alpha\tau_2]A_2'(\tau_2) - 1} \, . \tag{3.29}$$

3.1.3 Areas of convergence

Using the results of $U_1(z_1)$ and $U_2(z_2)$, we can give trivial but nevertheless useful results for the regions of convergence of $U(z_1, 0)$, $U(0, z_2)$ and $U(z_1, z_2)$. It holds that

$$U(z_1, 0) \text{ is analytic for all complex } z_1 \text{ with } |z_1| < \tau_1 \, , \tag{3.30}$$

and

$$U(0, z_2) \text{ is analytic for all complex } z_2 \text{ with } |z_2| < \tau_2 \, . \tag{3.31}$$

Furthermore, the region of convergence of the joint PGF $U(z_1, z_2)$ contains at least the region

$$\{(z_1, z_2) : |z_1| < \tau_1, |z_2| \leq 1\} \cup \{(z_1, z_2) : |z_1| \leq 1, |z_2| < \tau_2\} \, . \tag{3.32}$$

The proofs of these three statements are omitted, since they are exactly the same as those of Section 2.3.2.

3.2 Sufficient conditions for a geometric tail behavior

In the introduction of this chapter, we already imposed an assumption on the marginal arrival PGFs $A_1(z)$ and $A_2(z)$, namely Assumption 3.1. The following two conditions are obtained in this chapter, such that the joint pmf $p(i, j)$ has a geometric tail behavior:

Condition 3.1. *Let τ_1 and τ_2 be the unique real positive zeros, greater than one, of $K(z,1)$ and $K(1,z)$, respectively. Then,*

(a) $K^{(1)}(1,\tau_2) > 0$,

(b) $K^{(2)}(\tau_1,1) > 0$.

In terms of $A(z_1,z_2)$ these two conditions are written as

$$(a)\ \ \tau_2 - (1-\alpha)A_2(\tau_2) - (1-\alpha+\alpha\tau_2)A^{(1)}(1,\tau_2) > 0\,, \tag{3.33}$$

$$(b)\ \ \tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1+\alpha)A^{(2)}(\tau_1,1) > 0\,. \tag{3.34}$$

We first discuss some examples of arrival processes where these conditions are (not) fulfilled.

Condition 3.1 is always fulfilled when the type-1 and type-2 arrivals are uncorrelated, i.e. $A(z_1,z_2) = A_1(z_1)A_2(z_2)$. Indeed, in this case we have that

$$\begin{aligned} K^{(1)}(1,\tau_2) &= \tau_2 - (1-\alpha)A_2(\tau_2) - (1-\alpha+\alpha\tau_2)\lambda_1 A_2(\tau_2) \\ &= \tau_2(1-\lambda_1) - (1-\alpha)A_2(\tau_2) \\ &> \tau_2(\alpha-\lambda_1) \\ &> 0\,. \end{aligned} \tag{3.35}$$

Here we used (3.27) in the second step, the fact that $A_2(\tau_2) < \tau_2$ (which follows from (3.27)) in the third step and part of the stability condition in the last step. Likewise, we have that

$$\begin{aligned} K^{(2)}(\tau_1,1) &= \tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1+\alpha)A_1(\tau_1)\lambda_2 \\ &= \tau_1(1-\lambda_2) - \alpha A_1(\tau_1) \\ &> \tau_1(1-\alpha-\lambda_2) \\ &> 0\,, \end{aligned} \tag{3.36}$$

in which we used (3.24) in the second step, the fact that $A_1(\tau_1) < \tau_1$ (which follows from (3.24)) in the third step and part of the stability condition in the last step.

Next, we emphasize that Condition 3.1 *can* be true for dependent arrivals as well. For example, Condition 3.1 is fulfilled when the arrivals come from an output of an $N \times N$ queueing switch [17], i.e. the joint PGF of the arrivals within a time slot is given by

$$A(z_1,z_2) = \left(1 - \frac{\lambda_1}{N}(1-z_1) - \frac{\lambda_2}{N}(1-z_2)\right)^N, \quad N \in \mathbb{N}\,.$$

Indeed, since for this joint PGF

$$A^{(2)}(\tau_1,1) = \frac{\lambda_2 A(\tau_1,1)}{1-\lambda_1/N+(\lambda_1/N)\tau_1} < \lambda_2 A(\tau_1,1) = A_1(\tau_1)\lambda_2,$$

we can use the same inequalities that we used to prove (3.35).

However, it is possible to construct PGFs $A(z_1, z_2)$ such that Condition 3.1 does not hold. To demonstrate this, let us consider arrivals a_1, a_2, whose joint PGF is given by

$$A(z_1, z_2) = 1 - \frac{0.25}{5} + \frac{0.25}{5} z_1^5 z_2^5 \,. \tag{3.37}$$

It can be verified that

$$\begin{aligned} \lambda_1 &= 0.25, \\ \lambda_2 &= 0.25, \\ \mathrm{cov}[a_1, a_2] &= 1.1876. \end{aligned}$$

If we choose $\alpha = 0.5$, we have that $\tau_1 = \tau_2 = 1.283$ and

$$K^{(2)}(\tau_1, 1) = K^{(1)}(1, \tau_2) = -0.272 \,.$$

For the example above, it holds that $\mathrm{cov}[a_1, a_2] > 0$. Initially, our conjecture was that arrivals with negative covariance satisfy Condition 3.1 and arrivals with positive covariance do not satisfy Condition 3.1. However, these two conjectures turned out to be false. We now show counter-examples for both conjectures.

First, consider arrivals a_1, a_2 whose joint PGF is given by

$$A(z_1, z_2) = 0.65 + 0.05 z_2 + 0.3 z_1 z_2 \,. \tag{3.38}$$

It can be verified that

$$\begin{aligned} \lambda_1 &= 0.3, \\ \lambda_2 &= 0.35, \\ \mathrm{cov}[a_1, a_2] &= 0.195. \end{aligned}$$

If we choose $\alpha = 0.4$, we have that $\tau_1 = 1.555$, $\tau_2 = 2.785$ and

$$K^{(2)}(\tau_1, 1) = 0.4, \quad K^{(1)}(1, \tau_2) = 0.38 \,.$$

In our second counter-example, we show that there exist arrival processes with negative covariance such that Condition 3.1 is violated. Consider arrivals a_1, a_2 whose joint PGF is given by

$$A(z_1, z_2) = 0.655 + 0.1635 z_1 + 0.125 z_2 + 0.055 z_2^2 + 0.0015 z_1^5 z_2^5 \,. \tag{3.39}$$

It can be verified that the mean arrival rates are

$$\begin{aligned} \lambda_1 &= 0.1710, \\ \lambda_2 &= 0.2425, \end{aligned}$$

$$\mathrm{cov}[a_1, a_2] = -0.0039.$$

Let $\alpha = 0.5$, then it follows that $\tau_1 = 2.617$, $\tau_2 = 1.987$ and

$$K^{(2)}(\tau_1, 1) = -0.197.$$

The idea in the construction of this counterexample was to include a small probability that there can occur a large number of arrivals in both queues, but that the covariance is still negative.

Our initial conjecture turned out to be incorrect. We thus cannot make a precise statement in case of correlated arrivals. In the following sections, we will show that Condition 3.1 is sufficient for the (partial) generating functions $P_{1,n}(z)$ and $P_{2,n}(z)$ to have a simple pole at τ_1 and τ_2, respectively. Using Theorem 3.1, this will give rise to a geometric tail behavior of $p(i,j)$.

3.3 Asymptotic analysis of $U(z,0)$ and $U(0,z)$

In this section we show that the dominant singularity of $U(z,0)$ is τ_1 and that it is also a simple pole. Likewise, we show that the dominant singularity of $U(0,z)$ is τ_2 and that it is a simple pole as well. We start this section with the dominant singularity of $U(z,0)$.

Theorem 3.4. *Under Condition 3.1 (b), the function $U(z,0)$ has a simple pole at $z = \tau_1$. Moreover, the residue at $z = \tau_1$ of $U(z,0)$ equals*

$$\begin{aligned}&\operatorname*{res}_{z=\tau_1} U(z,0)\\ &= \frac{(\tau_1 - 1)(\alpha - \lambda_1)(\tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1 + \alpha)A^{(2)}(\tau_1, 1))}{(1-\alpha)\tau_1(1 - (1-\alpha)A_1(\tau_1) - ((1-\alpha)\tau_1 + \alpha)A_1'(\tau_1))}.\end{aligned} \tag{3.40}$$

Proof. We have that $K(z_1, z_2)$, as defined in (3.11), is bivariate analytic near $z_1 = \tau_1$, $z_2 = 1$. By the definition of τ_1, we have that $K(\tau_1, 1) = 0$. Moreover, we have that $K^{(2)}(\tau_1, 1) > 0$ because of Condition 3.1. By the implicit function theorem for analytic functions [11, Theorem B.4], a unique function $Y(z)$ and a radius $r > 0$ exist such that

1. $Y(z)$ is analytic in a neighbourhood $\{z \in \mathbb{C} : |z - \tau_1| < r\}$ of τ_1,
2. $Y(\tau_1) = 1$,
3. $K(z, Y(z)) = 0$ for $z \in \{z \in \mathbb{C} : |z - \tau_1| < r\}$.

Furthermore, we have that

$$Y'(\tau_1) = -\frac{K^{(1)}(\tau_1, 1)}{K^{(2)}(\tau_1, 1)}$$

$$= -\frac{1-(1-\alpha)A_1(\tau_1) - ((1-\alpha)\tau_1+\alpha)A_1'(\tau_1)}{\tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1+\alpha)A^{(2)}(\tau_1,1)} \,. \tag{3.41}$$

Moreover, since $K^{(1)}(\tau_1,1) < 0$ (see Lemma 3.1 on page 73) and $K^{(2)}(\tau_1,1) > 0$ (see Condition 3.1), we have that $Y'(\tau_1) > 0$. Consequently, we have that

$$Y(z) < 1, \quad \text{if } z \in]\tau_1 - \delta, \tau_1[\,,$$

for sufficiently small $\delta > 0$. From the discussions of Section 3.1.3, we know that $U(z_1,z_2)$ is bounded if $|z_1| < \tau_1$, $|z_2| \leq 1$. In particular, we have that $U(z,Y(z))$ is bounded for $z \in]\tau_1 - \delta, \tau_1[$. Hence, the right-hand side of the functional equation (3.10) vanishes for $\{z_1 = z, z_2 = Y(z)\}$, if $z \in]\tau_1 - \delta, \tau_1[$. We thus obtain that

$$(1-\alpha)z(Y(z)-1)U(z,0) + \alpha(z-1)Y(z)U(0,Y(z)) = 0, \quad z \in]\tau_1 - \delta, \tau_1[\,,$$

or

$$(1-\alpha)z(Y(z)-1)U(z,0) = -\alpha(z-1)Y(z)U(0,Y(z)), \quad z \in]\tau_1 - \delta, \tau_1[\,. \tag{3.42}$$

Both sides of this equation are analytic functions for $z \in]\tau_1 - \delta, \tau_1[$, because $U(z,0)$, $Y(z)$ and $U(0,Y(z))$ are analytic in this interval. However, we can take any sufficiently small $R > 0$, such that $|Y(z)| < \tau_2$ in $\{z \in \mathbb{C} : |z - \tau_1| < R\}$. Consequently, the RHS in (3.42) is an analytic function for $|z - \tau_1| < R$. Hence, we can analytically continue $(1-\alpha)z(Y(z)-1)U(z,0)$ into the region $|z - \tau_1| < R$ via (3.42). Since $Y(z)$ is analytic in this region, it follows that $U(z,0)$ is meromorphic in $|z - \tau_1| < R$. The poles of $U(z,0)$ in $|z - \tau_1| < R$ are the zeros of $Y(z) - 1$ (if any).

Because Y is analytic in τ_1 and $Y'(\tau_1) \neq 0$, it follows that Y is an injective function in a neighborhood of τ_1. Hence, we take any $R' \leq R$ such that Y is injective in $|z - \tau_1| < R'$. The only zero of $Y(z) - 1$ in this latter region is the point $z = \tau_1$. We conclude that τ_1 is a pole of $U(z,0)$.

We now prove that τ_1 is a simple pole of $U(z,0)$. Let us rewrite (3.42) as

$$U(z,0) = -\frac{\alpha(z-1)Y(z)U(0,Y(z))}{(1-\alpha)(Y(z)-1)z} \,.$$

Multiplying the equation above by $(z - \tau_1)$ and taking the limit to τ_1 yields

$$\begin{aligned}
&\lim_{z\to\tau_1} (z-\tau_1)U(z,0) \\
&= \lim_{z\to\tau_1} -\frac{(z-\tau_1)\alpha(z-1)Y(z)U(0,Y(z))}{(1-\alpha)(Y(z)-1)z} \\
&= -\frac{(\tau_1-1)(\alpha-\lambda_1)}{(1-\alpha)\tau_1 Y'(\tau_1)} \qquad (3.43) \\
&= \frac{(\tau_1-1)(\alpha-\lambda_1)(\tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1+\alpha)A^{(2)}(\tau_1,1))}{(1-\alpha)\tau_1(1-(1-\alpha)A_1(\tau_1) - ((1-\alpha)\tau_1+\alpha)A_1'(\tau_1))} \qquad (3.44)
\end{aligned}$$

Since $Y'(\tau_1) > 0$, $\tau_1 > 1$ and $\lambda_1 < \alpha$ (because of the stability condition), expression (3.43) is strictly negative. We can thus conclude that τ_1 is a simple pole of $U(z,0)$. □

We have proven that τ_1 is a simple pole of $U(z,0)$. From Subsection 3.1.3, we know that $U(z,0)$ is analytic for all complex z with $|z| < \tau_1$. Combining these two properties imply that τ_1 is a dominant singularity of $U(z,0)$. At this point we have not proven that there are no other singularities of $U(z,0)$ for $|z| = \tau_1$. Notice that there are no other zeros of the kernel $K(z,1)$ for $|z| = \tau_1$, see Lemma 3.1 on page 73. Relying on the implicit function theorem yet again, it can be proven that $U(z,0)$ is analytic in $|z| = \tau_1$, $z \neq \tau_1$. Intuitively, one also does not expect to find multiple singularities. If $U(z,0)$ would possess multiple singularities on the circle with radius τ_1, this would lead to periodicity of the coefficients $p(n,0)$ [11, IV. 6.1], which is very unlikely for the queueing model under consideration.

In a similar way as Theorem 3.4, we can prove the following theorem.

Theorem 3.5. *Under Condition 3.1 (a), the function $U(0,z)$ has a simple pole at $z = \tau_2$. Moreover, the residue at $z = \tau_2$ of $U(0,z)$ equals*

$$\begin{aligned} &\operatorname*{res}_{z=\tau_2} U(0,z) \\ &= \frac{(\tau_2 - 1)(1 - \alpha - \lambda_2)(\tau_2 - (1-\alpha)A_2(\tau_2) - (1 - \alpha + \alpha\tau_2)A^{(1)}(1,\tau_2))}{\alpha\tau_2(1 - \alpha A_2(\tau_2) - (1 - \alpha + \alpha\tau_2)A_2'(\tau_2))}. \end{aligned} \tag{3.45}$$

Finally, it can be proven that τ_2 is the dominant singularity of $U(0,z)$.

We conclude this section with an important remark with respect to Condition 3.1. In view of Theorem 3.4, one might wonder if it is not sufficient to assume that $K^{(2)}(\tau_1,1) \neq 0$ instead of assuming that $K^{(2)}(\tau_1,1) > 0$. Similarly, is it not sufficient to assume that $K^{(1)}(1,\tau_2) \neq 0$ instead of assuming that $K^{(1)}(1,\tau_2) > 0$? Although we explicitly needed Conditions 3.1 (a) and (b) in order to analytically continue the functions $U(z,0)$ and $U(0,z)$, suppose a *contrario* that $K^{(2)}(\tau_1,1) < 0$ and $K^{(1)}(1,\tau_2) < 0$. By taking the limits $\lim_{z\to\tau_1}(z-\tau_1)U(z,0)$ and $\lim_{z\to\tau_2}(z-\tau_2)U(0,z)$ it is seen that both limits are strictly positive. By virtue of Theorem 3.1, this yields that the coefficients of $U(z,0)$ and $U(0,z)$ are negative, a contradiction. Hence, τ_1 cannot be a singularity of $U(z_1,0)$ if $K^{(2)}(\tau_1,1) < 0$.

3.4 Asymptotic analysis of $P_{1,n}(z)$ and $P_{2,n}(z)$

In this section we obtain asymptotics for $p(i,j)$ for either i or j fixed. Our approach is that of determining the dominant singularity of the generating functions of $p(i,j)$, for fixed i or j. These generating functions are defined by $P_{1,n}(z)$

and $P_{2,n}(z)$, see (3.12) and (3.13). Recall once again that $P_{1,0}(z) = U(z,0)$ and that $P_{2,0}(z) = U(0,z)$. We thus already have obtained the dominant singularity of $P_{1,0}(z)$ and $P_{2,0}(z)$ in the previous section. Using the recurrence relations (3.19) and (3.20), we first show that τ_1 and τ_2 are *isolated* singularities of all $P_{1,n}(z)$ and $P_{2,n}(z)$, respectively, for $n = 1, 2, \ldots$. A point $\hat{z}$ is said to be an isolated singularity of a complex-valued function f if f is singular at $\hat{z}$ yet analytic in some deleted neighborhood of $\hat{z}$ [96, Definition 9.6].

Lemma 3.3. *τ_1 is an isolated singularity of $P_{1,n}(z)$, $n = 0, 1, 2, \ldots$*

Proof. Notice that

$$\begin{aligned} p(i,n) &\leq \sum_{j=0}^{\infty} p(i,j) \\ &= p_1(i) \; . \end{aligned}$$

It easily follows that the radius of convergence of $U_1(z)$ is a lower bound for the radius of convergence of $P_{1,n}(z)$. Consequently, the radius of convergence, and thus also the dominant singularity, of $P_{1,n}(z)$ is at least τ_1. Following a similar reasoning, it can be proven that the radius of convergence of $A_1(z)$ is a lower bound for the radius of convergence of $A_{1,n}(z)$. Hence, $A_{1,n}(z)$ is analytic for $|z| < \mathcal{R}_1$.

We have shown in Theorem 3.4 that τ_1 is a simple pole of $P_{1,0}(z)$. We prove by induction that τ_1 is an isolated singularity of $P_{1,n}(z)$, $\forall n \in \mathbb{N}$. Suppose that this is true for $n = 0, \ldots, m$. Then, by considering the Equation (3.19) for $n = m$ and solving for $P_{1,m+1}(z)$ it follows that

$$\begin{aligned} P_{1,m+1}(z) = \Bigg(& zP_{1,m}(z) - \alpha \sum_{j=0}^{m} A_{1,m-j}(z)[P_{1,j}(z) + (z-1)P_{1,j}(0)] \\ & -(1-\alpha)zA_{1,m}(z)P_{1,0}(z) - (1-\alpha)z \sum_{j=0}^{m-1} A_{1,m-j}(z)P_{1,j+1}(z) \Bigg) \; . \\ & \times \frac{1}{(1-\alpha)zA_{1,0}(z)} \; . \end{aligned} \tag{3.46}$$

From the expression above, we conclude that the possible singularities for $P_{1,m+1}(z)$ are the zeros of $(1-\alpha)zA_{1,0}(z)$, the singularities of $A_{1,m-j}(z)$ and the singularities of $P_{1,j}(z)$, $j = 0, \ldots, m$. Notice that the numerator of the above equation vanishes for $z = 0$. Furthermore, $A_{1,0}(z)$ has no positive real zeros in $[0, \mathcal{R}_1[$, since it is a partial PGF. Because of the induction hypothesis, τ_1 is an isolated singularity of the functions $P_{1,j}(z)$, $j = 0, \ldots, m$. Hence, in view of Equation (3.46), τ_1 is an isolated singularity of $P_{1,m+1}(z)$. □

Remark that according to Lemma 3.3, it is still possible that τ_1 is a removable singularity of $P_{1,n}(z)$. From equation (3.19) with $n = 0$, it is however easy to prove that τ_1 is a simple pole of $P_{1,1}(z)$. However, this is not easy to prove for general n. For example, let us assume that τ_1 is a simple pole of $P_{1,n}(z)$ for $n = 0, \ldots, m$. In view of equation (3.46), we see that the $\lim_{z\to\tau_1}(z-\tau_1)P_{1,m+1}(z)$ is a linear combination of the residues of $P_{1,j}(z)$ at τ_1. Hence, we do a priori not know that

$$\lim_{z\to\tau_1}(z-\tau_1)P_{1,m+1} \stackrel{?}{\neq} 0 \, .$$

Via a somewhat different reasoning, we will now show that the inequality above is indeed true.

Let us define the following sequence

$$B_n \triangleq \lim_{z\to\tau_1}(\tau_1 - z)P_{1,n}(z) \, . \tag{3.47}$$

Note that from (3.40),

$$B_0 = -\frac{(\tau_1-1)(\alpha-\lambda_1)(\tau_1-\alpha A_1(\tau_1)-((1-\alpha)\tau_1+\alpha)A^{(2)}(\tau_1,1))}{(1-\alpha)\tau_1(1-(1-\alpha)A_1(\tau_1)-((1-\alpha)\tau_1+\alpha)A_1'(\tau_1))} \, , \tag{3.48}$$

and that $B_0 > 0$, cf. (3.43).

Multiplying Equation (3.19) by $(\tau_1 - z)$ and taking the limit $z \to \tau_1$, we obtain that

$$\begin{aligned} \tau_1 B_n = \alpha \sum_{j=0}^{n} A_{1,n-j}(\tau_1)B_j + (1-\alpha)\tau_1 A_{1,n}(\tau_1)B_0 \\ + (1-\alpha)\tau_1 \sum_{j=0}^{n} A_{1,n-j}(\tau_1)B_{j+1}, \quad n = 0,1,2,\ldots \end{aligned} \tag{3.49}$$

Hence, we can recursively compute B_{n+1} in terms of $B_0, B_1, \ldots, B_n$ by the above equation, with B_0 given by (3.48). In Theorem 3.4, we have proven that $U(z,0)$ has a simple pole at $z = \tau_1$. Substituting $n = 0$ into (3.49) and solving for B_1 yields

$$B_1 = \frac{(\tau_1 - (\alpha + (1-\alpha)\tau_1)A_{1,0}(\tau_1))B_0}{(1-\alpha)\tau_1 A_{1,0}(\tau_1)}.$$

Because

$$(\alpha + (1-\alpha)\tau_1)A_{1,0}(\tau_1) < (\alpha + (1-\alpha)\tau_1)A_1(\tau_1) = \tau_1 \, ,$$

it follows that $B_1 > 0$. To prove that $B_n > 0$ for every n, we introduce the generating function of the sequence $\{B_n\}_{n=0}^{\infty}$. Let us denote this generating function by $B(z)$, i.e.

$$B(z) \triangleq \sum_{n=0}^{\infty} B_n z^n \, .$$

Multiplying all terms in (3.49) by z^n and summing over all valid n leads to the following expression for $B(z)$:

$$B(z) = \frac{(1-\alpha)\tau_1 A(\tau_1, z)(z-1)B_0}{\tau_1 z - ((1-\alpha)\tau_1 + \alpha z)A(\tau_1, z)} \ . \tag{3.50}$$

Using l'Hôpital's rule and Equation (3.48), it follows that $B(1) = C_1$, where C_1 was defined in (3.26). This result is not surprising since

$$\begin{aligned} C_1 &= \lim_{z\to\tau_1} (\tau_1 - z)U_1(z) \\ &= \lim_{z\to\tau_1} (\tau_1 - z) \sum_{n=0}^{\infty} P_{1,n}(z) \\ &= \sum_{n=0}^{\infty} \lim_{z\to\tau_1} (\tau_1 - z)P_{1,n}(z) \\ &= \sum_{n=0}^{\infty} B_n \\ &= B(1) \ , \end{aligned}$$

where we could switch limit and summation because the series converges uniformly for $|z| < \tau_1$.

Consider now the normalized function $\frac{B(z)}{C_1}$. We rewrite this function as follows

$$\begin{aligned} \frac{B(z)}{C_1} &= \frac{(1-\alpha)\tau_1 B_0}{C_1} \frac{A(\tau_1, z)(z-1)}{\tau_1 z - ((1-\alpha)\tau_1 + \alpha z)A(\tau_1, z)} \\ &= \frac{(1-\alpha)B_0}{C_1} \frac{A(\tau_1, z)(z-1)}{z - (1-\alpha+\frac{\alpha}{\tau_1}z)A(\tau_1, z)} \\ &= \frac{(1-\alpha)A_1(\tau_1)B_0}{C_1} \frac{\frac{A(\tau_1,z)}{A_1(\tau_1)}(z-1)}{z - ((1-\alpha)A_1(\tau_1) + \frac{\alpha A_1(\tau_1)}{\tau_1}z)\frac{A(\tau_1,z)}{A_1(\tau_1)}} \ . \end{aligned} \tag{3.51}$$

For ease of notation, let us define

$$\sigma \triangleq (1-\alpha)A_1(\tau_1) \ , \tag{3.52}$$

and

$$E(z) \triangleq \frac{A(\tau_1, z)}{A_1(\tau_1)} \ . \tag{3.53}$$

Since,

$$\tau_1 = ((1-\alpha)\tau_1 + \alpha)A_1(\tau_1) \ ,$$

see (3.24), we have that

$$1 = (1-\alpha)A_1(\tau_1) + \frac{\alpha A_1(\tau_1)}{\tau_1} \ . \tag{3.54}$$

Consequently,

$$\sigma < 1 \tag{3.55}$$

and

$$\frac{\alpha A_1(\tau_1)}{\tau_1} = 1 - \sigma \; . \tag{3.56}$$

Remark that $E(z)$ is a PGF. Indeed, using the definition of $A(z_1, z_2)$ we obtain that

$$E(z) = \sum_{j=0}^{\infty} \left(\frac{\sum_{i=0}^{\infty} a(i,j)\tau_1^i}{A_1(\tau_1)} \right) z^j \; .$$

From this expression, it is seen that the power series coefficients are positive. Since these coefficients sum up to one, $E(z)$ is a proper PGF.

Furthermore, from the expressions (3.48) and (3.26) we obtain that

$$\begin{aligned} \frac{B_0}{C_1} &= \frac{\tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1 + \alpha)A^{(2)}(\tau_1, 1)}{(1-\alpha)\tau_1 A_1(\tau_1)} \\ &= \frac{1 - (1-\sigma) - (1 - \alpha + \frac{\alpha}{\tau_1})A^{(2)}(\tau_1, 1)}{\sigma} \\ &= \frac{\sigma - \frac{A^{(2)}(\tau_1,1)}{A_1(\tau_1)}}{\sigma} \\ &= \frac{\sigma - E'(1)}{\sigma} \; , \end{aligned} \tag{3.57}$$

where in the third equality we used (3.54). Using (3.57), we can rewrite (3.51) as

$$\frac{B(z)}{C_1} = \frac{(\sigma - E'(1))(z-1)E(z)}{z - E(z)(\sigma + (1-\sigma)z)} \; . \tag{3.58}$$

Expression (3.58) is the PGF of the number of customers in a discrete-time Bernoulli model, cf. Section 1.5, with the probability σ that the server is available during a slot and with arrival PGF $E(z)$.

The stability condition of this queueing model is

$$E'(1) < \sigma \; ,$$

or

$$\frac{A^{(2)}(\tau_1, 1)}{A_1(\tau_1)} < (1-\alpha)A_1(\tau_1) \; .$$

This condition is equivalent to Condition 3.1 (b) on page 77. Indeed, the latter condition yields, cf. (3.34)

$$A^{(2)}(\tau_1, 1)((1-\alpha)\tau_1 + \alpha) < \tau_1 - \alpha A_1(\tau_1) \; .$$

The LHS and the RHS of the inequality above are equal to, cf. (3.24)

$$\begin{aligned} A^{(2)}(\tau_1,1)((1-\alpha)\tau_1+\alpha) &= A^{(2)}(\tau_1,1)\frac{\tau_1}{A_1(\tau_1)} \\ &= E'(1)\tau_1 \end{aligned}$$

and

$$\begin{aligned} \tau_1 - \alpha A_1(\tau_1) &= (1-\alpha)A_1(\tau_1)\tau_1 \\ &= \sigma\tau_1 \ , \end{aligned}$$

respectively. Hence, the stability condition of the Bernoulli model with the probability σ that the server is available during a slot and with arrival PGF $E(z)$ is equivalent with Condition 3.1 (b).

Since $\frac{B(z)}{C_1}$ can be defined as the PGF of the number of customers in a Bernoulli model, the coefficients B_n of $B(z)$ satisfy

$$B_n > 0, \quad \text{for every } n \ . \tag{3.59}$$

As a consequence, τ_1 is a simple pole of $P_{1,n}(z)$ for every n and it is the unique dominant singularity. To conclude, we obtain the following theorem.

Theorem 3.6. *The asymptotics for the joint probabilities $p(i,j)$ for large i are given by*

$$p(i,j) \sim B_j \tau_1^{-(i+1)}, \quad j = 0,1,\ldots \tag{3.60}$$

The coefficients B_j are recursively defined by (3.49) with initial condition (3.48).

In the first part of this section, we have focused on the partial PGFs $P_{1,n}(z)$. With a very similar analysis, we can obtain results for the partial PGFs $P_{2,n}(z)$. Define

$$D_n \triangleq \lim_{z\to\tau_2} (\tau_2 - z)P_{2,n}(z) \ . \tag{3.61}$$

It can be shown that

$$D_0 = -\frac{(\tau_2-1)(1-\alpha-\lambda_2)(\tau_2-(1-\alpha)A_2(\tau_2)-(1-\alpha+\alpha\tau_2)A^{(1)}(1,\tau_2))}{\alpha\tau_2(1-\alpha A_2(\tau_2)-(1-\alpha+(1-\alpha)\tau_2)A_2'(\tau_2))} \ , \tag{3.62}$$

and

$$\begin{aligned} \tau_2 D_n &= (1-\alpha)\sum_{j=0}^{n} A_{2,n-j}(\tau_2)D_j + \alpha\tau_2 A_{2,n}(\tau_2)D_0 \\ &\quad + \alpha\tau_2 \sum_{j=0}^{n} A_{2,n-j}(\tau_2)D_{j+1}, \quad n = 0,1,2,\ldots \end{aligned} \tag{3.63}$$

The generating function of the sequence $\{D_n\}_n^\infty$ is obtained as

$$\begin{aligned} D(z) &\triangleq \sum_{n=0}^{\infty} D_n z^n \\ &= \frac{\alpha\tau_2 A(z,\tau_2)(z-1)D_0}{\tau_2 z - (1-\alpha+\alpha\tau_2)A(z,\tau_2)} \,. \end{aligned} \tag{3.64}$$

It can be shown that $\frac{D(z)}{C_2}$ is the PGF of the number of customers of a discrete-time Bernoulli model, in steady-state. The stability condition of this model is equivalent to Condition 3.1 (a). The theorem analogous to Theorem 3.6 is then given by

Theorem 3.7. *The asymptotics for the joint probabilities $p(i,j)$ for large j are given by*

$$p(i,j) \sim D_i \tau_2^{-(j+1)}, \quad i = 0, 1, \ldots \tag{3.65}$$

The coefficients D_j are recursively defined by (3.63) with initial condition (3.62).

3.5 A more detailed analysis: independent arrivals in the two queues

In Section 3.2, we showed that Condition 3.1 is naturally fulfilled in case of independent arrivals. In this section, we will show that much more details about the regions of convergence of $U(z,0)$ and $U(0,z)$ can be uncovered in this particular case. We commence this section with a more detailed analysis of the kernel K.

3.5.1 Analysis of the kernel K

In this subsection, we will examine some key properties of $K(z_1,z_2)$ with respect to the values τ_1 and τ_2 in the case that $A(z_1,z_2) = A_1(z_1)A_2(z_2)$. The results obtained in this subsection can be seen as generalizations of Lemma 2.2 on page 32 and Lemma 2.4 on page 34.

We first investigate the function $z_1 \mapsto K(z_1,\tau_2)$ for real values of z_1. As a reminder:

$$K(z,\tau_2) = z\tau_2 - ((1-\alpha)z + \alpha\tau_2)A_1(z)A_2(\tau_2) \,,$$

where τ_2 is defined in Lemma 3.2 on page 75. We have the following lemma.

Lemma 3.4.

1. *Equation $K(z_1,\tau_2) = 0$ has exactly two real positive roots inside the interval $[0,\mathcal{R}_1[$, 1 and say ω_1, such that $1 < \omega_1 < \mathcal{R}_1$.*

2. $K^{(1)}(\omega_1, \tau_2) < 0$.

3. $K(z_1, \tau_2) = 0$ *has no other roots with the same absolute value as* ω_1.

Proof. (1) and (2): We have that

$$K^{(11)}(z_1, \tau_2) = -2(1-\alpha)A_1'(z_1)A_2(\tau_2) - ((1-\alpha)z_1 + \alpha\tau_2)A_1''(z_1)A_2(\tau_2) \; .$$

Because $A_1(z)$ is a PGF, it follows that $A_1(z) > 0$, $A_1'(z) > 0$ and $A_1''(z) > 0$ for $z \in [0, \mathcal{R}_1[$. For the same reason, we have that $A_2(\tau_2) > 0$. We can conclude that

$$K^{(11)}(z_1, \tau_2) < 0 \quad \text{for } z_1 \in [0, \mathcal{R}_1[\; .$$

By the definition of τ_2, we have that $z_1 = 1$ is a root of $K(z_1, \tau_2) = 0$. Moreover, $K^{(1)}(1, \tau_2) > 0$ (see (3.35)). Finally, $K(z_1, \tau_2) \to -\infty$ as $z_1 \to \mathcal{R}_1$ because of Assumption 3.1. Combining these observations yields the result.

(3): Consider z_1 such that $|z_1| = \omega_1$, $z_1 \neq \omega_1$. Since

$$|(1-\alpha)z_1 + \alpha\tau_2| < (1-\alpha)\omega_1 + \alpha\tau_2 \; ,$$

we have that

$$\begin{aligned} |((1-\alpha)z_1 + \alpha\tau_2)A_1(z_1)A_2(\tau_2)| &< ((1-\alpha)|z_1| + \alpha\tau_2)A_1(|z_1|)A_2(\tau_2) \\ &= ((1-\alpha)\omega_1 + \alpha\tau_2)A_1(\omega_1)A_2(\tau_2) \\ &= \omega_1 \\ &= |z_1| \; . \end{aligned}$$

Hence, (3) is proven. □

In other words, the graph of the function $K(x, \tau_2)$ resembles that of a parabola that opens downward and which intersects the x axis at $x = 1$ and $x = \omega_1$ (with $\omega_1 > 1$), see Figure 3.2.

Let us now consider the function $K(\tau_1, z_2)$ for real values of z_2. Note that

$$K(\tau_1, z) = \tau_1 z - ((1-\alpha)\tau_1 + \alpha z)A(\tau_1, z) \; .$$

We mention the following equivalent of Lemma 3.4.

Lemma 3.5.

1. *Equation* $K(\tau_1, z_2) = 0$ *has exactly two real positive roots inside the interval* $\in [0, \mathcal{R}_2[$, 1 *and say* ω_2, *such that* $1 < \omega_2 < \mathcal{R}_2$.

2. $K^{(2)}(\tau_1, \omega_2) < 0$.

3. $K(\tau_1, z_2) = 0$ *has no other roots with the same absolute value as* ω_2.

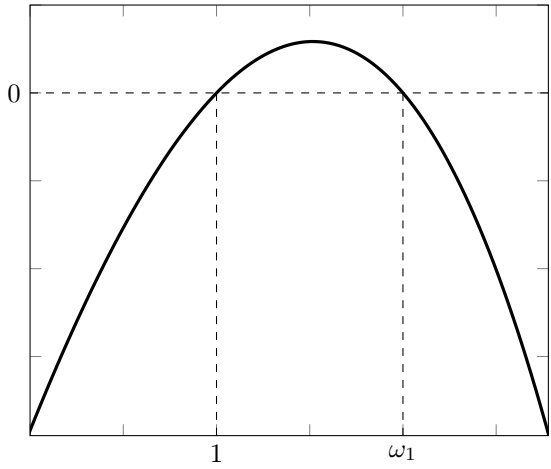


Figure 3.2: Illustration of the function $K(x, \tau_2)$ for positive real x.

We can specify the location of τ_1 and τ_2 relative to ω_1 and ω_2, respectively. We have the following result.

Lemma 3.6. *The following inequalities are valid*

$$1 < \tau_1 < \omega_1 < \mathcal{R}_1 \tag{3.66}$$

and

$$1 < \tau_2 < \omega_2 < \mathcal{R}_2 \; . \tag{3.67}$$

Proof. We only have to prove that $\tau_1 < \omega_1$ and $\tau_2 < \omega_2$. From Lemma 3.4 and Lemma 3.5 it follows that the functions $z_2 \mapsto K(\tau_1, z_2)$ and $z_1 \mapsto K(z_1, \tau_2)$ are strictly positive when $z_2 \in]1, \omega_2[$ and $z_1 \in]1, \omega_1[$, respectively. Furthermore, it also follows that the functions $z_2 \mapsto K(\tau_1, z_2)$ and $z_1 \mapsto K(z_1, \tau_2)$ are strictly negative when $z_2 \in]\omega_2, \mathcal{R}_2[$ and $z_1 \in]\omega_1, \mathcal{R}_1[$, respectively. Hence, it is sufficient to show that $K(\tau_1, \tau_2) > 0$.

Using (3.24) and (3.27), we can write

$$\tau_1 \tau_2 = ((1-\alpha)\tau_1 + \alpha) A_1(\tau_1)(1 - \alpha + \alpha\tau_2) A_2(\tau_2) \; .$$

If we substitute the expression above into the expression of $K(\tau_1, \tau_2)$, we obtain that

$$\begin{aligned} K(\tau_1, \tau_2) &= \tau_1\tau_2 - ((1-\alpha)\tau_1 + \alpha\tau_2) A_1(\tau_1) A_2(\tau_2) \\ &= ((1-\alpha)\tau_1 + \alpha) A_1(\tau_1)(1 - \alpha + \alpha\tau_2) A_2(\tau_2) \\ &\quad - ((1-\alpha)\tau_1 + \alpha\tau_2) A_1(\tau_1) A_2(\tau_2) \\ &= [(1-\alpha)\tau_1 + \alpha)(1 - \alpha + \alpha\tau_2) - (1-\alpha)\tau_1 - \alpha] A_1(\tau_1) A_2(\tau_2) \end{aligned}$$

$$= \alpha(1-\alpha)(\tau_1 - 1)(\tau_2 - 1)A_1(\tau_1)A_2(\tau_2)$$
$$> 0 .$$

□

We will now couple the values z_1 and z_2 such that $K(z_1, z_2) = 0$. The following three technical lemmas will be used often in the next subsections.

Lemma 3.7. *For every z_2 for which $1 < |z_2| < \omega_2$, there exists a unique zero, say $X(z_2)$, such that $K(X(z_2), z_2) = 0$ and $|X(z_2)| < \tau_1$. Additionally, if $1 < |z_2| < \tau_2$, it holds that $|X(z_2)| < 1$.*

Proof. First, by virtue of Lemma 3.5, we have that

$$\tau_1 x - ((1-\alpha)\tau_1 + \alpha x)A_1(\tau_1)A_2(x) = K(\tau_1, x) > 0, \quad x \in]1, \omega_2[. \tag{3.68}$$

Secondly, let the complex value z_2 be fixed, $1 < |z_2| < \omega_2$. On $|z_1| = \tau_1$, we have

$$\begin{aligned} |((1-\alpha)z_1 + \alpha z_2)A_1(z_1)A_2(z_2)| &\leq ((1-\alpha)|z_1| + \alpha|z_2|)A_1(|z_1|)A_2(|z_2|) \\ &= ((1-\alpha)\tau_1 + \alpha|z_2|)A(\tau_1)A_2(|z_2|) . \end{aligned}$$

On the other hand, we have that $|z_1 z_2| = \tau_1|z_2|$. Because of (3.68), we have the inequality

$$|((1-\alpha)z_1 + \alpha z_2)A_1(z_1)A_2(z_2)| < \tau_1|z_2| .$$

By virtue of Rouché's theorem, the number of zeros of $z_1 \mapsto z_1 z_2$ inside $|z_1| < \tau_1$ is then the same as the number of zeros of $z_1 \mapsto K(z_1, z_2)$. The former number is 1 (due to the trivial zero $z_1 = 0$). Hence, we have found that for fixed z_2, $1 < |z_2| < \omega_2$, the function $z_1 \mapsto K(z_1, z_2)$ has exactly one zero inside the disk $|z_1| = \tau_1$, say $X(z_2)$.

Notice now that, by virtue of Lemma 3.2 (on page 75),

$$x - (1 - \alpha + \alpha x)A_2(x) = K(1, x) > 0, \quad x \in]1, \tau_2[. \tag{3.69}$$

Let the complex value z_2 be fixed, $1 < |z_2| < \tau_2$. On $|z_1| = 1$, we have

$$\begin{aligned} |((1-\alpha)z_1 + \alpha z_2)A_1(z_1)A_2(z_2)| &\leq ((1-\alpha)|z_1| + \alpha|z_2|)A_1(|z_1|)A_2(|z_2|) \\ &= (1 - \alpha + \alpha|z_2|)A_2(|z_2|) . \end{aligned}$$

On the other hand, we have that $|z_1 z_2| = |z_2|$. Because of (3.69), we have the inequality

$$|((1-\alpha)z_1 + \alpha z_2)A_1(z_1)A_2(z_2)| < |z_2| .$$

Application of Rouché's theorem yields that for fixed z_2, $1 < |z_2| < \tau_2$, the function $z_1 \mapsto K(z_1, z_2)$ has exactly one zero inside the disk $|z_1| = 1$. Because of uniqueness, it necessarily follows that this zero is $X(z_2)$. □

Lemma 3.8. *For every z_1 for which $1 < |z_1| < \omega_1$, there exists a unique zero, say $Y(z_1)$, such that $K(z_1, Y(z_1)) = 0$ and $|Y(z_1)| < \tau_2$. Additionally, if $1 < |z_1| < \tau_1$, it holds that $|Y(z_1)| < 1$.*

Proof. The proof is similar to that of Lemma 3.7 and is therefore omitted. □

Lemma 3.9. *$X(z)$ and $Y(z)$ are analytic functions for $z \in \{z \in \mathbb{C} : 1 < |z| < \omega_2\}$ and $z \in \{z \in \mathbb{C} : 1 < |z| < \omega_1\}$, respectively.*

Proof. The function $K(z_1, z_2)$ is bivariate analytic at tuples (z_1, z_2) such that $|z_1| < \mathcal{R}_1$ and $|z_2| < \mathcal{R}_2$. In particular, $K(z_1, z_2)$ is bivariate analytic at tuples (z_1, z_2) such that $|z_1| < \tau_1$ and $1 < |z_2| < \omega_2$. Consider a complex value $\hat{z}_2$ such that $1 < |\hat{z}_2| < \omega_2$. Since $X(\hat{z}_2)$ is the *unique* zero of $z_1 \mapsto K(z_1, \hat{z}_2)$ we have that $K(X(\hat{z}_2), \hat{z}_2) = 0$ and $K^{(1)}(X(\hat{z}_2), \hat{z}_2) \neq 0$. Consequently, the implicit function theorem for analytic functions implies that $X(z_2)$ is analytic at $\hat{z}_2$.

The proof for $Y(z)$ is analogous and is therefore omitted. □

Finally, we present two more technical lemmas that will be used at the very end of this chapter.

Lemma 3.10. *If $K^{(2)}(\omega_1, \tau_2) > 0$, then the function $Y(z)$ can be analytically continued in a neighborhood of ω_1, such that $Y(\omega_1) = \tau_2$ and $Y'(\omega_1) > 0$.*

Proof. We have that $K(z_1, z_2)$, as defined in (3.11), is jointly analytic near $z_1 = \omega_1$, $z_2 = \tau_2$. By the definition of ω_1, we further have that $K(\omega_1, \tau_2) = 0$.

By the implicit function theorem for analytic functions, a unique function $\check{Y}(z)$ and a radius $r > 0$ exist such that

1. $\check{Y}(z)$ is analytic in a neighbourhood V_{ω_1} of ω_1,
2. $\check{Y}(\omega_1) = \tau_2$,
3. $K(z, \check{Y}(z)) = 0$ for $z \in V_{\omega_1}$.

Further, we have that

$$\check{Y}'(\omega_1) = -\frac{K^{(1)}(\omega_1, \tau_2)}{K^{(2)}(\omega_1, \tau_2)} \,. \tag{3.70}$$

We know from Lemma 3.4 on page 87 that $K^{(1)}(\omega_1, \tau_2) < 0$. Because we assume that $K^{(2)}(\omega_1, \tau_2) > 0$, we obviously have that $\check{Y}'(\omega_1) > 0$.

Because $\check{Y}'(\omega_1) > 0$, we have that $\check{Y}(z) < \tau_2$, $z \in]\omega_1 - \delta, \omega_1[$, for sufficiently small $\delta > 0$. In view of Lemma 3.8, $Y(z)$ is defined as the *unique* function such

that $K(z, Y(z)) = 0$, $|Y(z)| < \tau_2$ for z, $1 < |z| < \omega_1$. But we also have that $K(z, \check{Y}(z)) = 0$, $\check{Y}(z) < \tau_2$ for $z \in]\omega_1 - \delta, \omega_1[$. Hence, it must be that

$$Y(z) = \check{Y}(z), \quad z \in]\omega_1 - \delta, \omega_1[\, . \tag{3.71}$$

It follows that $\check{Y}(z)$ is the unique analytic continuation of $Y(z)$ in V_{ω_1}. □

Lemma 3.11. *If $K^{(1)}(\tau_1, \omega_2) > 0$, then the function $X(z)$ can be analytically continued in a neighborhood of ω_2, such that $X(\omega_2) = \tau_1$ and $X'(\omega_2) > 0$.*

Proof. The proof is similar to that of Lemma 3.10 and is therefore omitted. □

3.5.2 Refinement for the singularity analysis of $U(z, 0)$ and $U(0, z)$

Using the results from the previous section, we are now ready to present a series of results concerning the analytic behavior of $U(z, 0)$ and $U(0, z)$. The final results are summarized at the end of this section. We first show that $U(z, 0)$ and $U(0, z)$ can be meromorphically continued outside the open unit disk into larger disks.

Theorem 3.8. *$U(z, 0)$ has a meromorphic continuation to the annulus $\tau_1 \leq |z| < \omega_1$. The poles of $U(z, 0)$ in $\tau_1 \leq |z| < \omega_1$ (if any) are the zeros of $(Y(z) - 1)$.*

Proof. Because of Lemma 3.8, tuples $(z, Y(z))$ such that $1 < |z| < \tau_1$ belong to the set defined in (3.32). Therefore, substituting $\{z_1 = z, z_2 = Y(z)\}$ with $1 < |z| < \tau_1$ into the fundamental functional equation (3.10) yields

$$(1 - \alpha)(Y(z) - 1)zU(z, 0) + \alpha(z - 1)Y(z)U(0, Y(z)) = 0 \, ,$$

or

$$(1 - \alpha)(Y(z) - 1)zU(z, 0) = -\alpha(z - 1)Y(z)U(0, Y(z)) = 0 \, . \tag{3.72}$$

Both sides of the equation above are analytic functions for z, $1 < |z| < \tau_2$.

In view of Lemma 3.8, we have that $|Y(z)| < \tau_2$ when $1 < |z| < \omega_1$. Using (3.31) and Lemma 3.9, we obtain that the RHS of (3.72) represents an analytic function for $1 < |z| < \omega_1$. This analytic function agrees with the LHS for $1 < |z| < \tau_1$. Hence, we can analytically continue the LHS to $\tau_1 \leq |z| < \omega_1$ via (3.72). Because $(Y(z) - 1)zU(z, 0)$ is analytic in $\tau_1 \leq |z| < \omega_1$, it follows that $U(z, 0)$ is meromorphic in $\tau_1 \leq |z| < \omega_1$. The poles of $U(z, 0)$ inside $\tau_1 \leq |z| < \omega_1$ (if any) are the zeros of $(Y(z) - 1)$. □

We mention the following equivalent of Theorem 3.8, but with respect to function $U(0, z)$.

Theorem 3.9. *$U(0,z)$ has a meromorphic continuation to the annulus $\tau_2 \leq |z| < \omega_2$. The poles of $U(0,z)$ in $\tau_2 \leq |z| < \omega_2$ (if any) are the zeros of $(X(z)-1)$.*

Having determined regions in which $U(z,0)$ and $U(0,z)$ are meromorphic, we now have to determine the poles of these two functions inside these regions. Using Theorem 3.4 on page 79, it follows that τ_1 is the unique dominant singularity of $U(z,0)$. It is natural to ask if $U(z,0)$ has other singularities in the area $\tau_1 < |z| < \omega_1$. In view of Theorem 3.8, the candidate singularities are zeros of $Y(z)-1$. This means that if $K(q,1)=0$ with $\tau_1 < |q| < \omega_1$, then q is a singularity of $U(z,0)$. We emphasize that the existence of such zeros is possible. For example, consider arrivals a_1, a_2 whose PGF is given by

$$A_1(z_1) = 1 - \frac{\lambda_1}{2} + \frac{\lambda_1}{2} z^2 \,,$$

and

$$A_2(z_2) = \left(1 - \frac{\lambda_2}{7} + \frac{\lambda_2}{7} z_2\right)^7 \,,$$

respectively. If we choose the following system parameters

$$\lambda_1 = 0.25, \quad \lambda_2 = 0.15, \quad \alpha = 0.4 \,,$$

we have that

$$\tau_1 = 1.4820, \quad \tau_2 = 5.809, \quad \omega_1 = 3.312 \,.$$

The function $K(z_1,1)$ is a third degree polynomial with zeros 1, τ_1 and $q^* := -3.1487$. Hence,

$$|q^*| < \omega_1 \quad \text{and} \quad Y(q^*) = 1 \,.$$

We can conclude that q^* is a pole of $U(z,0)$. Its residue can be computed as

$$\lim_{z \to q^*} (z - q^*) U(z,0) = -\frac{\alpha - \lambda_1}{1-\alpha} \left(1 - \frac{1}{q^*}\right) \frac{1}{Y'(q^*)} \,.$$

Furthermore, the quantity $Y'(q^*)$ can be evaluated as $Y'(q^*) = -\frac{K^{(1)}(q^*,1)}{K^{(2)}(q^*,1)}$.

A powerful technique to determine all the zeros of an analytic function in subregions of the complex plane is described by [97]. This method can be applied to the function $K(z,1)$ in the region $\tau_1 < |z| < \omega_1$. It suffices to apply the method to the upper-half of the annulus $\tau_1 < |z| < \omega_1$, i.e. $\mathrm{Im}(z) \geq 0$, because if $K(z,1)=0$, then also $K(\bar{z},1)=0$. Since these regions are compact, $K(z,1)$ can only have a finite number (possibly zero) of zeros inside $\tau_1 < |z| < \omega_1$. Hence, we have the following theorem.

Theorem 3.10. *Suppose there are L_1 zeros of $K(z,1)$, say $q_{1,k}$, $k = 1, \ldots, L_1$, such that $\tau_1 < |q_{1,k}| < \omega_1$. Then $q_{1,k}$ is a pole of $U(z,0)$.*

Proof. The proof is simple and is therefore omitted. □

We conclude the singularity analysis of $U(z,0)$ with one final question: can ω_1 be a singularity of $U(z,0)$? We provide the answer through the following theorem.

Theorem 3.11. *If $K^{(2)}(\omega_1,\tau_2) > 0$, then ω_1 is a simple pole of $U(z,0)$ with residue*

$$\operatorname*{res}_{z=\omega_1} U(z,0) = \frac{\alpha(\omega_1-1)\tau_2}{(1-\alpha)\omega_1(\tau_2-1)} \frac{K^{(2)}(\omega_1,\tau_2)}{K^{(1)}(\omega_1,\tau_2)} \operatorname*{res}_{z=\tau_2} U(0,z)\,. \tag{3.73}$$

The main idea for the proof of this theorem is as follows. Looking at Eq. (3.72), the RHS has a simple pole for $z=\omega_1$. This is because we have that $Y(\omega_1)=\tau_2$ (guaranteed by Lemma 3.10 on page 91), $Y'(\omega_1)\neq 0$ and τ_2 is a simple pole of $U(0,z_2)$. Therefore, ω_1 must also be a singularity of the LHS of Eq. (3.72). Because $Y(z)$ is analytic in ω_1, it is a singularity of $U(z,0)$. We will now prove this more rigorously.

Proof of Theorem 3.11. Consider again Eq. (3.72). We have already shown that this equation is also valid for $1<|z|<\omega_1$. Using Lemma 3.10 on page 91, it follows that Y is analytic in an open neighborhood V_{ω_1} of ω_1. Moreover, $Y(\omega_1)=\tau_2$ and $Y'(\omega_1)>0$. Hence, we can always find an open neighborhood V'_{ω_1} of ω_1 such that

1. $|Y(z)|<\omega_2$ if $z\in V'_{\omega_1}$,
2. $V'_{\omega_1}\cap\{z\in\mathbb{C}\,:\,1<|z|<\omega_1\}\neq\emptyset$.

By virtue of Lemma 3.9 on 93, $U(0,z)$ is an analytic function inside $|z_2|<\omega_2$, $z_2\neq\tau_2$. Therefore, the composition $U(0,Y(z))$ is analytic inside $z\in V'_{\omega_1}$, $z\neq\omega_1$. Hence, the RHS of Eq. (3.72) is an analytic function for $z\in V'_{\omega_1}$. It follows that we can analytically continue the LHS of (3.72) into $z\in V'_{\omega_1}$, $z\neq\omega_1$. In summary, we have proven that $(Y(z)-1)zU(z,0)$ is analytic in $z\in V'_{\omega_1}$, $z\neq\omega_1$.

From Eq. (3.72), it follows that $\lim_{z\to\omega_1}(Y(z)-1)zU(z,0)=\infty$. Hence ω_1 is a pole of $U(z,0)$. Using transformation of residues, we can compute the residue of $U(z,0)$ at $z=\omega_1$. We obtain that

$$\begin{aligned}
\lim_{z\to\omega_1}(z-\omega_1)U(z,0) &= -\frac{\alpha}{1-\alpha}\frac{\omega_1-1}{\omega_1}\frac{\tau_2}{\tau_2-1}\lim_{z\to\omega_1}(z-\omega_1)U(0,Y(z)) \\
&= -\frac{\alpha}{1-\alpha}\frac{\omega_1-1}{\omega_1}\frac{\tau_2}{\tau_2-1}\lim_{z\to\tau_2}(z-\tau_2)U(0,z)\frac{1}{Y'(\omega_1)} \\
&= \frac{\alpha(\omega_1-1)\tau_2}{(1-\alpha)\omega_1(\tau_2-1)}\ \frac{K^{(2)}(\omega_1,\tau_2)}{K^{(1)}(\omega_1,\tau_2)}\operatorname*{res}_{z=\tau_2}U(0,z)\,.
\end{aligned}$$

□

In summary, for the function $U(z,0)$ we have proven for the special case of $A(z_1,z_2) = A_1(z_1)A_2(z_2)$ that

1. $U(z,0)$ is analytic in $|z| < \tau_1$. The radius of convergence of $U(z,0)$ is τ_1.
2. $U(z,0)$ can be meromophically continued into $\tau_1 \leq |z| < \omega_1$.
3. τ_1 is always a simple pole of $U(z,0)$ and this is the singularity of $U(z,0)$ with the smallest norm.
4. $U(z,0)$ can have (if any) a finite number of additional poles in the region $\tau_1 < |z| < \omega_1$. These are the roots of $K(z,1) = 0$, if any.
5. If $K^{(2)}(\omega_1,\tau_2) > 0$, then ω_1 is a simple pole of $U(z,0)$.
6. All residues of the aforementioned poles can be computed exactly.

Without proof, we state the equivalent theorems of Theorem 3.10 and Theorem 3.11 with respect to the function $U(0,z)$.

Theorem 3.12. *Suppose there are L_2 zeros of $K(1,z)$, say $q_{2,k}$, $k = 1,\ldots,L_2$, such that $\tau_2 < |q_{2,k}| < \omega_2$. Then $q_{2,k}$ is a pole of $U(0,z)$.*

Theorem 3.13. *If $K^{(1)}(\tau_1,\omega_2) > 0$, then ω_2 is a simple pole of $U(0,z)$ with residue*

$$\operatorname*{res}_{z=\omega_2} U(0,z) = \frac{(1-\alpha)(\omega_2-1)\tau_1}{\alpha\omega_2(\tau_1-1)} \frac{K^{(1)}(\tau_1,\omega_2)}{K^{(2)}(\tau_1,\omega_2)} \operatorname*{res}_{z=\tau_1} U(z,0)\,. \tag{3.74}$$

In summary, for the function $U(0,z)$ it can be proven, for the special case of $A(z_1,z_2) = A_1(z_1)A_2(z_2)$, that

1. $U(0,z)$ is analytic in $|z| < \tau_2$. The radius of convergence of $U(0,z)$ is τ_2.
2. $U(0,z)$ can be meromophically continued into $\tau_2 \leq |z| < \omega_2$.
3. τ_2 is always a simple pole of $U(0,z)$ and this is the singularity of $U(0,z)$ with the smallest norm.
4. $U(0,z)$ can have (if any) a finite number of additional poles in the region $\tau_2 < |z| < \omega_2$. These are the roots of $K(1,z) = 0$, if any.
5. If $K^{(1)}(\tau_1,\omega_2) > 0$, then ω_2 is a simple pole of $U(0,z)$.
6. All residues of the aforementioned poles can be computed exactly.

3.5.3 Further discussion

To check if ω_1 is a singularity of $U(z,0)$, one can simply evaluate $K^{(2)}(\omega_1,\tau_2)$. Is it possible to roughly predict when $K^{(2)}(\omega_1,\tau_2)>0$ in terms of the value of τ_2? First and foremost, we emphasize that simple examples can be found such that $K^{(2)}(\omega_1,\tau_2)<0$. However, we will show that if τ_2 is *close to* 1, then it is likely that $K^{(2)}(\omega_1,\tau_2)>0$.

Consider $K^{(2)}(x,y)$ as a function in $\mathbb{R}^2$. It is clear that $K^{(2)}(x,y)$ is continuous in $(\tau_1,1)$. Hence, $K^{(2)}(x,y)$ does not change sign in a (real) neighborhood of $(\tau_1,1)$, i.e. $\exists\delta$, such that

$$K^{(2)}(x,y)>0 \quad \text{if } \sqrt{(x-\tau_1)^2+(y-1)^2}<\delta\ .$$

The tuple (ω_1,τ_2) belongs to this neighborhood if

$$\sqrt{(\omega_1-\tau_1)^2+(\tau_2-1)^2}<\delta\ . \tag{3.75}$$

Let $\tau_2=1+\varepsilon$, $\varepsilon>0$. Remark that ω_1 is defined as the unique zero of $z\mapsto K(z,\tau_2)$ in $]1,\mathcal{R}_1[$. To underline the dependency of $\tau_2=1+\varepsilon$, we write $\omega_1(\varepsilon)$. The inequality above becomes

$$\sqrt{(\omega_1(\varepsilon)-\tau_1)^2+\varepsilon^2}<\delta\ . \tag{3.76}$$

Intuitively, if τ_2 is close to 1 (i.e. ε is close to zero), then ω_1 will be close to τ_1. Moreover, $\omega_1(0)=\tau_1$ by definition of τ_1. Hence, the bound (3.76) will be satisfied if ε is small enough.

3.6 Concluding remarks

In this chapter, we focused on the asymptotic behavior of the joint pmf $p(i,j)$ of the system contents. We showed that requiring $K^{(2)}(\tau_1,1)>0$ and $K^{(1)}(1,\tau_2)>0$ is sufficient for obtaining a geometric asymptotic behavior of $p(i,j)$. We further showed that these intriguing conditions are equivalent to the stability condition of a related queueing model. Even with the successful completion of these results, several new questions arise. In this concluding section, we mention the research gaps we encountered throughout this chapter and did not have the time or inventiveness for.

Remark that at the end of Section 3.3 we showed that if $K^{(2)}(\tau_1,1)<0$, then τ_1 cannot be a singularity of $U(z,0)$. In order that $K^{(2)}(\tau_1,1)>0$ is a necessary condition for τ_1 to be a singularity of $U(z,0)$, we should also handle the boundary case $K^{(2)}(\tau_1,1)=0$. Hence, we have the following question.

Question 3.1. *What happens in the case that* $K^{(2)}(\tau_1,1)=0$ *and/or* $K^{(1)}(1,\tau_2)=0$*?*

We discussed in Section 3.2 some examples of arrival processes where the conditions $K^{(2)}(\tau_1, 1) > 0$ and/or $K^{(1)}(1, \tau_2) > 0$ are not fulfilled. Hence, we have the following natural question.

Question 3.2. *Can we determine the dominant singularity of* $U(z, 0)$ *and* $U(0, z)$ *in the case that* $K^{(2)}(\tau_1, 1) < 0$ *and/or* $K^{(1)}(1, \tau_2) < 0$*?*

Note that if (for example) $K^{(2)}(\tau_1, 1) < 0$, then the dominant singularity of $U(z, 0)$ is strictly greater than τ_1. The main hindrance for answering Question 3.2 is that we do not have a clue about the location of the dominant singularity in this case.

As mentioned before, we showed that $K^{(2)}(\tau_1, 1) > 0$ is equivalent to the stability condition of a related queueing model. It goes without saying that this striking result cannot be a coincidence. Therefore, one final research question we would have liked to study is the following.

Question 3.3. *Can we give an intuitive interpretation to the condition* $K^{(2)}(\tau_1, 1) > 0$ *(and* $K^{(1)}(1, \tau_2) > 0$*)? What is the connection of the related queueing model with the original queueing model?*

4

Approximate analysis: a novel approximation method

In this chapter, we shift focus back to the complete joint pmf $p(i,j)$ of the system contents. The pmf $p(i,j)$ is only known for a few special cases, see also Chapter 2. In the general case, we obtained in the previous chapter an intriguing condition (namely Condition 3.1 on page 77) in order to obtain asymptotic expressions for $p(i,j)$ when either $i \to +\infty$ or $j \to +\infty$. In this chapter, we explore how these asymptotic results can be exploited to obtain the probabilities $p(i,j)$ that are not in the tail. Results for the latter probabilities are not only interesting measures on their own, but are also required for the computation of performance measures like the variance of the total system content, the fraction of time that the server idles, etc. We make the same restrictions as in Chapter 3, namely Assumption (3.1) and Condition 3.1. For definiteness, we repeat these restrictions here. Let the radius of convergence of $A_j(z) = \sum_{n=0}^{\infty} a_j(n) z^n$ be denoted by $\mathcal{R}_j$ and define the τ_i as the unique solution of the equations below

$$\tau_1 = ((1-\alpha)\tau_1 + \alpha) A_1(\tau_1) \,, \tag{4.1}$$

$$\tau_2 = (1 - \alpha + \alpha\tau_2) A_2(\tau_2) \,, \tag{4.2}$$

with $1 < \tau_i < \mathcal{R}_i$.

Assumptions:

1. $\mathcal{R}_j > 1$,
2. $\lim_{z \to \mathcal{R}_j} A_j(z) = +\infty$,
3. $K^{(2)}(\tau_1, 1) > 0$,
4. $K^{(1)}(1, \tau_2) > 0$.

These assumptions are sufficient to obtain the asymptotic expressions of the previous chapter. We will use these asymptotic expressions in this chapter.

A key contribution of our paper [98] is a novel approximation method. Broadly speaking, we approximate the boundary functions $U(z_1, 0)$ and $U(0, z_2)$ in the functional equation (2.12) by rational functions. An approximation as a rational function is obtained in two steps. First the dominant singularities (and the corresponding residues) of $U(z_1, 0)$ and $U(0, z_2)$ are found. This step is actually the result of Chapter 3. This result is used to approximate the coefficients of the Taylor series of $U(z_1, 0)$ and $U(0, z_2)$, starting from a certain index onwards. In the second step, the remaining finite number of coefficients is approximated by substituting the Taylor series of the first step into the functional equation (2.12). Linear equations between this finite number of unknowns are found by considering zeros-tuples of the kernel $K(z_1, z_2)$, such that the right-hand side of the functional equation vanishes. This novel approximation method is also the key contribution of this chapter.

This chapter is organized as follows. In the following section we survey some of the most well-established approximation methods for multidimensional queueing models. In Section 4.2, we present our novel approximation method. The approximation results are validated against simulation results in Section 4.3. We discuss further insights in Section 4.4, before giving some concluding remarks in Section 4.5.

4.1 State-of-the-art and related approximation methods

Because of the drawbacks of the boundary-value-problem theory, some approximation techniques for two- as well as multidimensional queues have been investigated in the past. We briefly discuss the state-of-the-art.

A first prominent approach is the **Power Series Approximation Approach** (PSA) [18] for two-queue models. In [18], the joint PGF of the numbers of customers in both queues is expressed as a power series in a specified parameter of the model. The coefficients of the consecutive terms in this power series are calculated recursively (either numerically or analytically). Truncation or any other approximation based on the knowledge of a finite number of these coefficients are the result of this technique. The major drawback of the PSA approach as per [18] is that in practice only few series coefficients can be calculated within reasonable computing time [99]. The PSA was originally developed to directly approximate the joint pmf (instead of the joint PGF) in [100, 101]. The joint pmf is then expressed as a power series (usually in the load). The coefficients of the power series are calculated recursively via the equilibrium equations. The disadvantage of the original PSA approach is that it does not provide error bounds and that it suffers from the curse of dimensionality [102].

If we would mimic the approach of [18], we should write $U(z_1, z_2)$ as a power series expansion in α (instead of the load) and substitute this power series into (2.12). Unfortunately, in our case, this PSA approach as per [18] does not seem to be applicable since for the boundary values $\alpha = 0$ or $\alpha = 1$, one of the two queues becomes unstable (assuming that $\lambda_1 > 0$ and $\lambda_2 > 0$).

Series expansion techniques for Markov chains are sometimes referred to as **perturbation techniques**. Perturbation methods are mainly motivated by sensitivity analysis of the results with respect to some system parameter [103]. In particular singular perturbations where the perturbation does not preserve the class-structure of the non-perturbed Markov chain, have received considerable attention in literature [104]. Perturbation techniques can be used as a numerical solution technique for two-dimensional queueing models, see for example [38, 105] and references therein.

Among the class of nearest-neighbor random walks with no one-step displacements to the North, North-East and East, it can be shown that the boundary functions $U(z, 0)$ and $U(0, z)$ are meromorphic functions [78]. The so-called **compensation method** [106] is a renowned analytic-algorithmic method that can be considered as an alternative to analytic continuation when meromorphicity of the boundary functions is established. The compensation approach works directly in the probability domain and only works for the specific subclass of random walks as in [78]. Several extensions of the original compensation approach were developed over the years. For example, the studies [107–109] indicate that the compensation method can be extended to a class of so-called multi-layered two-dimensional random walks [107]. Furthermore, the compensation method was generalized for three-dimensional models under additional restrictions on the one-step displacements [110]. In [111] the compensation method was applied to a two-dimensional random walk on the lattice of the first quadrant with arbitrarily sized jumps on the boundaries and one-step displacements to the North (lower part of the first quadrant) or the East (upper part of the quadrant). Finally, it has been shown in [112, 113] that the framework of the compensation method can be extended to a two-dimensional random walk with bounded one-step displacements to non-neighboring states. Several restrictions still apply on the one-step displacements in all the aforementioned extensions. Therefore, the compensation method appears to be inapplicable to our queueing model with general $A(z_1, z_2)$.

Another, less satisfactory, approach is a brute-force **truncation of the state space**. Broadly speaking, we are then solving a finite-capacity model. A finite Markov chain can in principle always be analyzed. However, for large capacities (or more than two queues) this chain becomes large and solving it is time-consuming (the curse of dimensionality yet again). Instead of assuming that all queue capacities are finite, it is better to assume that all queues but one are finite. Such a system can be analyzed efficiently using the celebrated **matrix-geometric and matrix-analytic methods** [4, 5]. The effect of state-space truncation has received considerable attention, see e.g. [114–116]. In many of

these studies, it is shown that the tail behavior can differ between the finite and infinite capacity system, as the finite capacity grows. Extensions of the matrix-analytic approach to nearest-neighbor random walks in the positive quadrant exist. See for example [60, 117] and the references therein.

Instead of a state-space truncation or a perturbation in a system parameter, one can search for a (slightly) modified system for which there is an explicit characterization of the joint system-content distribution. This may be resulting in analytical error bounds for the performance of interest in the original model. Such bounds are provided by the **Markov reward approach** [118]. The Markov reward error bounds can be formulated as linear optimization problems [119, 120]. As noted in [120, Chapter 2], the linear programs for upper and lower bounds are not always feasible. In particular, once the load exceeds some threshold the problems often become infeasible and cannot return any bounds. Some intuitions to this problem are given in Chapter 7 of [120].

4.2 Approximation for $U(z,0)$ and $U(0,z)$

The subject of this section is to find an efficient approximation for the coefficients of the power series

$$U(z,0) = \sum_{n=0}^{\infty} p(n,0) z^n \,, \tag{4.3}$$

$$U(0,z) = \sum_{n=0}^{\infty} p(0,n) z^n \,. \tag{4.4}$$

In view of Theorem 3.6 (see page 86) and Theorem 3.7 (see page 87), we have that

$$p(n,0) \sim B_0 \tau_1^{-(n+1)} \,, \tag{4.5}$$

and

$$p(0,n) \sim D_0 \tau_2^{-(n+1)} \,. \tag{4.6}$$

For definiteness, we repeat the expressions for B_0 and D_0. We have that

$$B_0 = -\frac{(\tau_1 - 1)(\alpha - \lambda_1)(\tau_1 - \alpha A_1(\tau_1) - ((1-\alpha)\tau_1 + \alpha)A^{(2)}(\tau_1, 1))}{(1-\alpha)\tau_1(1 - (1-\alpha)A_1(\tau_1) - ((1-\alpha)\tau_1 + \alpha)A_1'(\tau_1))} \,, \tag{4.7}$$

$$D_0 = -\frac{(\tau_2 - 1)(1 - \alpha - \lambda_2)(\tau_2 - (1-\alpha)A_2(\tau_2) - (1 - \alpha + \alpha\tau_2)A^{(1)}(1, \tau_2))}{\alpha\tau_2(1 - \alpha A_2(\tau_2) - (1 - \alpha + (1-\alpha)\tau_2)A_2'(\tau_2))} \,. \tag{4.8}$$

Our purpose is to obtain accurate approximations for $p(n,0)$, $p(0,n)$, for every index n. Hence, for the indices for which the dominant-pole approximation is not sufficiently accurate, we have to approximate $p(n,0)$ and $p(0,n)$ in a

different way. To accomplish this, we show in the following lemma that the PGFs $U(z,0)$ and $U(0,z)$ can be approximated by simple functions.

Lemma 4.1. *Let r_1 be any number between τ_1 and the modulus of the next singularity of $U(z,0)$. Let r_2 be any number between τ_2 and the modulus of the next singularity of $U(0,z)$. We have that*

$$\left| U(z,0) - \left(\sum_{n=0}^{M} p(n,0) z^n + \frac{B_0}{\tau_1^{M+1}} \frac{z^{M+1}}{\tau_1 - z} \right) \right| < C \frac{r_1}{r_1 - \tau_1} \left(\frac{\tau_1}{r_1} \right)^{M+1}, \ |z| < \tau_1 , \tag{4.9}$$

$$\left| U(0,z) - \left(\sum_{n=0}^{M} p(0,n) z^n + \frac{D_0}{\tau_2^{M+1}} \frac{z^{M+1}}{\tau_2 - z} \right) \right| < \tilde{C} \frac{r_2}{r_2 - \tau_2} \left(\frac{\tau_2}{r_2} \right)^{M+1}, \ |z| < \tau_2 , \tag{4.10}$$

with C and $\tilde{C}$ positive constants.

Proof. By the residue theorem, we can write that

$$\frac{1}{2\pi \mathrm{i}} \int_{|z|=r_1} \frac{U(z,0)}{z^{n+1}} dz = p(n,0) - \frac{B_0}{\tau_1^{n+1}} , \tag{4.11}$$

where we denote with $|z| = r_1$ the positively oriented circle with radius r_1. The contour integral is $O(r_1^{-n})$ as $n \to \infty$, so that we obtain again the dominant-pole approximations (4.5) and (4.6).

Next, for ease of notation, we define the contour integrals as

$$I_n \triangleq \frac{1}{2\pi \mathrm{i}} \int_{|z|=r_1} \frac{U(z,0)}{z^{n+1}} dz , \tag{4.12}$$

$$J_n \triangleq \frac{1}{2\pi \mathrm{i}} \int_{|z|=r_2} \frac{U(0,z)}{z^{n+1}} dz , \tag{4.13}$$

such that we can compactly write

$$p(n,0) = \frac{B_0}{\tau_1^{n+1}} + I_n , \tag{4.14}$$

$$p(0,n) = \frac{D_0}{\tau_2^{n+1}} + J_n . \tag{4.15}$$

Substituting (4.14) for $n = M+1, M+2, \ldots$ into (4.3), yields

$$U(z,0) = \sum_{n=0}^{M} p(n,0) z^n + \frac{B_0}{\tau_1^{M+1}} \frac{z^{M+1}}{\tau_1 - z} + \sum_{n=M+1}^{\infty} I_n z^n, \quad |z| < \tau_1 . \tag{4.16}$$

Whence,

$$U(z,0) - \left(\sum_{n=0}^{M} p(n,0) z^n + \frac{B_0}{\tau_1^{M+1}} \frac{z^{M+1}}{\tau_1 - z} \right) = \sum_{n=M+1}^{\infty} I_n z^n, \quad |z| < \tau_1 \ .$$

Using that $|z| < \tau_1$ and $|I_n| \leq C r_1^{-n}$, for some constant $C > 0$, we obtain that

$$\left| \sum_{n=M+1}^{\infty} I_n z^n \right| \leq \sum_{n=M+1}^{\infty} |I_n z^n| < C \frac{r_1}{r_1 - \tau_1} \left(\frac{\tau_1}{r_1} \right)^{M+1} \ .$$

Hence, (4.9) is proven.

Analogously, we can obtain (4.10). □

In view of the lemma above, we approximate the functions $U(z,0)$ and $U(0,z)$ by the functions

$$U(z,0) \approx \hat{p}(0,0) + \sum_{n=1}^{M} \hat{p}(n,0) z^n + \frac{B_0}{\tau_1^{M+1}} \frac{z^{M+1}}{\tau_1 - z}, \quad |z| < \tau_1 \ , \tag{4.17}$$

$$U(0,z) \approx \hat{p}(0,0) + \sum_{n=1}^{M} \hat{p}(0,n) z^n + \frac{D_0}{\tau_2^{M+1}} \frac{z^{M+1}}{\tau_2 - z}, \quad |z| < \tau_2 \ , \tag{4.18}$$

where the $\hat{p}(n,0)$, $\hat{p}(0,n)$ are approximations for $p(n,0)$ and $p(0,n)$, respectively.

4.2.1 Estimation of the remaining probabilities

Since $U(z_1, z_2)$ is a joint PGF satisfying (2.12), see page 26, it must be that

$$(1-\alpha)(w_j - 1) v_j U(v_j, 0) + \alpha (v_j - 1) w_j U(0, w_j) = 0 \ , \tag{4.19}$$

for any finite set of tuples (v_j, w_j), $j = 0, \ldots, N$ such that $|v_j| \leq 1$, $|w_j| \leq 1$ and $K(v_j, w_j) = 0$. We will now substitute Equation (4.16) (with $z = v_j$) into the equation above. Moreover, if we similarly substitute

$$U(0, w_j) = \sum_{n=0}^{M} p(0,n) w_j^n + \frac{D_0}{\tau_2^{M+1}} \frac{w_j^{M+1}}{\tau_2 - w_j} + \sum_{n=M+1}^{\infty} J_n w_j^n \ ,$$

we obtain that

$$\begin{aligned} &(1-\alpha)(w_j - 1) v_j \sum_{n=0}^{M} p(n,0) {v_j}^n + \alpha (v_j - 1) w_j \sum_{n=0}^{M} p(0,n) {w_j}^n \\ &= -(1-\alpha)(w_j - 1) \frac{B_0 {v_j}^{M+2}}{\tau_1^{M+1}(\tau_1 - v_j)} - \alpha (v_j - 1) \frac{D_0 {w_j}^{M+2}}{\tau_2^{M+1}(\tau_2 - w_j)} \\ &\quad - (1-\alpha)(w_j - 1) v_j \sum_{n=M+1}^{\infty} I_n {v_j}^n - \alpha (v_j - 1) w_j \sum_{n=M+1}^{\infty} J_n {w_j}^n \ . \end{aligned} \tag{4.20}$$

If we introduce the vectors $\boldsymbol{\varepsilon}$ and $\mathbf{s}$ with components $(j = 0, \ldots, N)$

$$\varepsilon_j = (1-\alpha)(w_j - 1)v_j \sum_{n=M+1}^{\infty} I_n {v_j}^n + \alpha(v_j - 1)w_j \sum_{n=M+1}^{\infty} J_n {w_j}^n \,, \tag{4.21}$$

$$s_j = -(1-\alpha)(w_j - 1)\frac{B_0 {v_j}^{M+2}}{\tau_1^{M+1}(\tau_1 - v_j)} - \alpha(v_j - 1)\frac{D_0 {w_j}^{M+2}}{\tau_2^{M+1}(\tau_2 - w_j)} \,, \tag{4.22}$$

we can rewrite Equation (4.20) as

$$(1-\alpha)(w_j - 1)v_j \sum_{n=0}^{M} p(n,0){v_j}^n + \alpha(v_j - 1)w_j \sum_{n=0}^{M} p(0,n){w_j}^n = s_j - \varepsilon_j \,.$$

Or in matrix notation

$$T\mathbf{p} + \boldsymbol{\varepsilon} = \mathbf{s} \,, \tag{4.23}$$

where $\mathbf{p}$ is the vector of unknown probabilities $p(0,0), \ldots, p(M,0), p(0,1), \ldots, p(0,M)$, $\mathbf{s}$ is a known vector and T is the known coefficient matrix with entries

$$T_{j,k} = \begin{cases} (1-\alpha)(w_j - 1)v_j + \alpha(v_j - 1)w_j & k = 0 \,, \\ (1-\alpha)(w_j - 1)v_j^{k+1}, & 1 \le k \le M \,, \\ \alpha(v_j - 1)w_j^{k+1-M}, & M+1 \le k \le 2M \,. \end{cases} \tag{4.24}$$

The equations (4.23) can be seen as observations s_i that are noise-perturbed representations of a linear transformation of $\mathbf{p}$. Notice that we can make as many observations N as we want. If we choose $N = 2M$, i.e. there are as many observations as unknowns, we can estimate $\mathbf{p}$ by $\hat{\mathbf{p}}$

$$\hat{\mathbf{p}} = T^{-1}\mathbf{s} \,, \tag{4.25}$$

assuming that T is invertible and $\boldsymbol{\varepsilon}$ is small. The error that we make is $\hat{\mathbf{p}} - \mathbf{p} = T^{-1}\boldsymbol{\varepsilon}$.

The function $U(z_1, z_2)$ can then be approximated by replacing $U(z_1, 0)$ and $U(0, z_2)$ by (4.17) and (4.18) in expression (2.16).

The only thing left to do, is to decide *which* zeros (v_j, w_j) we will use in our approximation method.

4.2.2 A suitable set of zero-tuples

There are infinitely many possible choices for the zeros of K, while for our approximation method we only need a finite set. First of all, it is important that the zeros can easily be found numerically. Hence, information about the location of the zeros is desired. Secondly, the entries (4.24) of the coefficient matrix T are also determined by the zeros and therefore the error $T^{-1}\varepsilon$ depends on the choice of the zeros. Since it is a priori not clear what is the better choice, we take inspiration from the boundary-value approach. The boundary-value approach relies on the following lemma.

Lemma 4.2. *Suppose S_1 and S_2, with $S_1 \subset \{z_1 : |z_1| \leq 1\}$ and $S_2 \subset \{z_2 : |z_2| \leq 1\}$, are simple, smooth and non-self intersecting contours, and there exists a one-to-one map $g : S_2 \mapsto S_1$, such that for every $z_2 \in S_2$, $(g(z_2), z_2)$ is a zero-tuple of the kernel K.*
If functions $U(z,0)$ and $U(0,z)$, both analytic in $|z| < 1$, satisfy (4.19) for (z_1, z_2), with $z_1 = g(z_2)$, $z_2 \in S_2$, then $U(z,0)$ and $U(0,z)$ satisfy (4.19) for all (z_1, z_2) with $|z_1| \leq 1$, $|z_2| \leq 1$.

Proof. The proof is given in [121, Sect. 1]. □

If we thus could find simple, smooth and non-self intersecting contours S_1 and S_2, such that $(1-\alpha)(z_2-1)z_1U(z_1,0) + \alpha(z_1-1)z_2U(0,z_2) = 0$ for all $z_1 \in S_1$, $z_2 \in S_2$, then by analytic continuation $(1-\alpha)(z_2-1)z_1U(z_1,0) + \alpha(z_1-1)z_2U(0,z_2) = 0$ for all (z_1, z_2) with $|z_1| \leq 1$, $|z_2| \leq 1$. It is therefore reasonable to choose a finite set of zeros from such particular contours.

As in [121, Sect. 6], we can consider tuples of the form $(ze^{\mathrm{i}\varphi}, ze^{-\mathrm{i}\varphi})$, with $\varphi \in [0, 2\pi[$. The equation $K(ze^{\mathrm{i}\varphi}, ze^{-\mathrm{i}\varphi}) = 0$ is equivalent with

$$z^2 = [(1-\alpha)ze^{\mathrm{i}\varphi} + \alpha ze^{-\mathrm{i}\varphi}]A(ze^{\mathrm{i}\varphi}, ze^{-\mathrm{i}\varphi}) \, .$$

Canceling the common factor z on both sides of the equation yields

$$z = ((1-\alpha)e^{\mathrm{i}\varphi} + \alpha e^{-\mathrm{i}\varphi})A(ze^{\mathrm{i}\varphi}, ze^{-\mathrm{i}\varphi}) \, . \tag{4.26}$$

A direct application of Rouché's theorem implies that

Theorem 4.1. *Equation (4.26) has, for fixed φ, exactly one zero $z =: f(e^{\mathrm{i}\varphi})$ satisfying $|z| < 1$. Further we have that $\lim_{\varphi \to 0} f(e^{\mathrm{i}\varphi}) = 1$.*

Proof. See [49, Pg. 188]. □

The sets

$$S_1 = \{f(e^{i\varphi})e^{i\varphi} \mid \varphi \in [0, 2\pi[\} \, , \tag{4.27}$$

$$S_2 = \{f(e^{i\varphi})e^{-i\varphi} \mid \varphi \in [0, 2\pi[\} \tag{4.28}$$

present a pair of contours which satisfy the conditions discussed above [121].

The discussion above has led to the following choice of zeros

$$v_j = f(e^{\mathrm{i}\varphi_j})e^{\mathrm{i}\varphi_j}, \quad w_j = f(e^{\mathrm{i}\varphi_j})e^{-\mathrm{i}\varphi_j} \, .$$

We point out some key properties of the choice of zeros. First, multiplying both sides of Equation (4.26) by minus one and using the fact that $-1 = e^{\mathrm{i}\pi}$, yields

$$f(e^{\mathrm{i}(\varphi+\pi)}) = -f(e^{\mathrm{i}\varphi}) \, ,$$

whence,

$$f(e^{\mathrm{i}(\varphi+\pi)})e^{\mathrm{i}(\varphi+\pi)} = f(e^{\mathrm{i}\varphi})e^{\mathrm{i}\varphi} \quad \text{and} \quad f(e^{\mathrm{i}(\varphi+\pi)})e^{-\mathrm{i}(\varphi+\pi)} = f(e^{\mathrm{i}\varphi})e^{-\mathrm{i}\varphi} \; .$$

The contours S_1 and S_2 are therefore traversed twice if φ goes from 0 to 2π, such that φ can be restricted to $[0,\pi[$.

Secondly, we exploit the fact that if (z_1, z_2) is a zero-tuple of K, then the complex conjugate $(\overline{z_1}, \overline{z_2})$ is a zero-tuple of K as well. Therefore, an additional linear equation is found by using the complex conjugate zero-tuple $(\overline{v_j}, \overline{w_j})$. This will force $\hat{\mathbf{p}}$ to be real. Moreover, this halves the number of zero-tuples that have to be found and we can further restrict φ to $[0,\pi/2[$. By restricting φ to this interval, first only that part of S_1 and S_2 that lies in the upper-half plane is considered. But by also adding the complex conjugates, values in the lower-half plane are also considered.

We sample M equidistant values φ_j, $j = 0,\ldots,M-1$ in $]0,\pi/2[$ and compute $f(e^{\mathrm{i}\varphi_j})$ by means of a root-finding algorithm (e.g. the Newton-Raphson method). An additional number of M linear equations is found by using the complex conjugate zeros $(\overline{v_j}, \overline{w_j})$. The final equation that can be used is the normalization condition $(1-\alpha)U(1,0)+\alpha U(0,1) = 1-\lambda_1-\lambda_2$. This equation is obtained by considering the PGF of the total number of customers, i.e.

$$U(z,z) = \frac{A(z,z)(z-1)((1-\alpha)U(z,0)+\alpha U(0,z))}{z-A(z,z)} \; .$$

If we take the limit $z \to 1$ in the expression above and use l'Hôpital's rule, we obtain that $(1-\alpha)U(1,0)+\alpha U(0,1) = 1-\lambda_1-\lambda_2$. The approximated equation reads

$$(1-\alpha)\sum_{n=0}^{M}\hat{p}(n,0)+\alpha\sum_{n=0}^{M}\hat{p}(0,n) = 1-\lambda_1-\lambda_2-\frac{(1-\alpha)B_0}{\tau_1^{M+1}(\tau_1-1)}-\frac{\alpha D_0}{\tau_2^{M+1}(\tau_2-1)} \; .$$

The above linear equation in $\hat{p}(0,0),\ldots,\hat{p}(M,0),\hat{p}(0,1),\ldots,\hat{p}(0,M)$ yields an additional row to the coefficient matrix T and an additional entry to the column vector $\mathbf{s}$:

$$T_{2M,k} = \begin{cases} 1, & k = 0\,; \\ 1-\alpha, & 1 \le k \le M; \\ \alpha, & M+1 \le k \le 2M \; , \end{cases} \tag{4.29}$$

$$s_{2M} = 1-\lambda_1-\lambda_2-\frac{(1-\alpha)B_1}{\tau_1^{M+1}(\tau_1-1)}-\frac{\alpha B_2}{\tau_2^{M+1}(\tau_2-1)} \; . \tag{4.30}$$

Summarizing, our numerical approach consists of the following steps:

1. Compute τ_1 and τ_2 (which boils down to computing a zero of a nonlinear equation in a single variable, see (4.1) and (4.2)).

2. Compute B_0 and D_0 using (4.7) and (4.8), respectively.
3. Solve a relatively small linear system $T\hat{\mathbf{p}} = \mathbf{s}$. The entries $T_{j,k}$ of T are given by (4.29) and (4.24) for $j = 0, \ldots, 2M-1$ with

$$\begin{aligned} v_j &= f(e^{\mathrm{i}\varphi_j})e^{\mathrm{i}\varphi_j}, \\ w_j &= f(e^{\mathrm{i}\varphi_j})e^{-\mathrm{i}\varphi_j}, \\ v_{M+j} &= \overline{v_j}, \\ w_{M+j} &= \overline{w_j}, \end{aligned} \qquad j = 0, \ldots, M-1\,. \tag{4.31}$$

   The entries s_j of $\mathbf{s}$ are given by (4.30) and (4.22) for $j = 0, \ldots, 2M-1$.

In order to compute the vectors $\mathbf{v}$ and $\mathbf{w}$, $f(e^{\mathrm{i}\varphi_j})$, $j = 0, \ldots, M-1$, is computed by computing the unique zero inside the unit disk of Equation (4.26).

The condition number of the matrix T relative to a norm $\|\cdot\|$ is defined as $\|T\| \cdot \|T^{-1}\|$. The condition number serves as a measure of stability for the linear system $T\hat{\mathbf{p}} = \mathbf{s}$. Since the columns in the matrix T are powers of the vectors v and w, the condition number of T is often too large (which indicates instability) for high values of M, resulting in a (almost) singular coefficient matrix. Therefore, in order to keep the linear system stable, we expect that the truncation parameter M needs to be small. On the other hand, the bounds in Lemma 4.1 require that M needs to be large enough. The strength of the approach is that in case of light to moderate load, the tail-probabilities are already close to the exact values for small M. We will also demonstrate this in Sect. 4.3.

Besides the choice of truncation parameter M, there are plenty of possible choices for the values φ_j, $j = 0, \ldots, M-1$. Therefore, the matrix T can be seen as function of φ_j, i.e. $T(\varphi_0, \varphi_1, \ldots, \varphi_{M-1})$. In our main experiments in Sect. 4.3, we have chosen φ_j equidistant in $]0, \pi/2[$, since this seems the most natural and is easy to implement. We will briefly come back to this choice in Sect. 4.3 as well.

4.3 Validation of the approximation method

In this section we compare the obtained approximations to simulation results. We assume that the customer types have different, mutually independent, arrival distributions. More specifically, we assume that the number of type-1 arrivals per slot are geometrically distributed with mean λ_1, i.e.

$$A_1(z) = \frac{1}{1 + \lambda_1 - \lambda_1 z}\,. \tag{4.32}$$

Further, we assume that the number of type-2 arrivals per slot are Poisson distributed with mean λ_2, i.e.

$$A_2(z) = e^{\lambda_2(z-1)}\,. \tag{4.33}$$

In Sect. 4.2 we have proposed a novel method to estimate the probabilities $p(n,0)$ and $p(0,n)$ for low index values of n. In order to investigate the accuracy of this approximation method, we selected 5 arbitrary cases $(\alpha, \lambda_1, \lambda_2)$ from the possible parameter space, the parameters of which are listed in Table 4.1. We additionally give τ_1, τ_2 and the total mean system content $\mathsf{E}[u_1 + u_2]$ to have a rough idea about the traffic regime. Notice that the total mean system content can be computed exactly by taking the sum of the marginal mean system contents. The parameters are generated in a manner such that the total arrival rate $\lambda_1 + \lambda_2$ is at least 0.3.

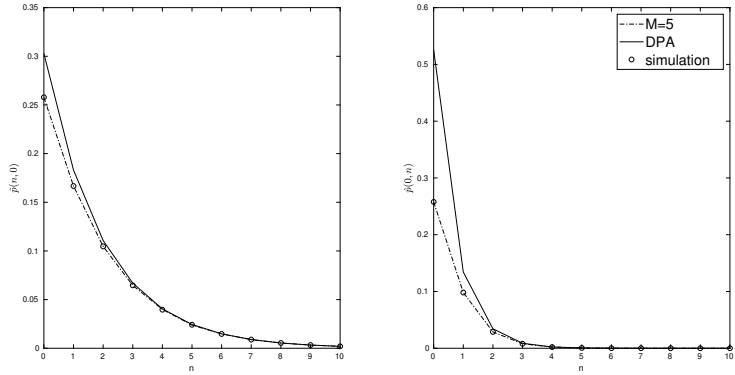
Table 4.1: Parameter settings for the 5 random cases to validate the accuracy of the approximation method to estimate $p(n,0)$ and $p(0,n)$ for low index values n.

Case	α	λ_1	λ_2	τ_1	τ_2	$\mathsf{E}[u_1 + u_2]$
1	0.587462	0.355125	0.125898	1.654240	3.914652	1.940062
2	0.478253	0.258507	0.396966	1.850058	1.429588	3.726257
3	0.796212	0.396695	0.177450	2.007114	1.165356	7.132573
4	0.617399	0.541399	0.288171	1.140377	1.407085	9.735655
5	0.214682	0.164818	0.762360	1.302539	1.049787	23.85433

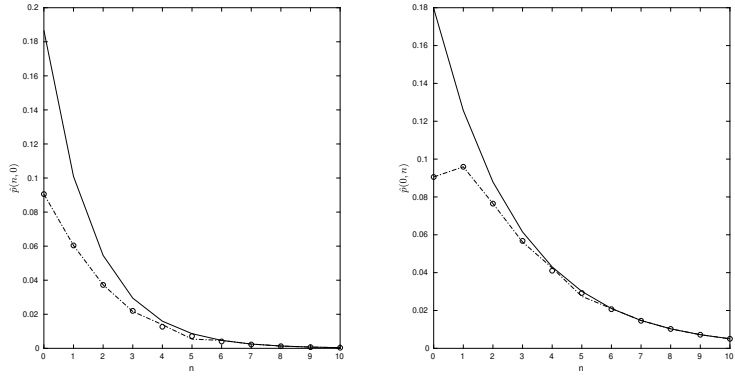
Each subplot of Figure 4.1 and Figure 4.2 shows the approximations $\hat{p}(n,0)$ (left) and $\hat{p}(0,n)$ (right) for $n = 0, 1, \ldots, 10$. The approximation method is applied for different values of M and with φ_j equidistantly chosen in $]0, \pi/2[$ (see also the discussion in Sect. 4.2.2). In addition, we also show the dominant-pole approximations (DPA) (4.5) and (4.6). We have simulated the queueing system (marks in the figure) for assessing the accuracy. The simulation results were obtained based on 10^9 slots.

Figure 4.1 (a) and (b) show that the approximations are very accurate for the particular parameter settings in Case 1 and Case 2. Notice that the dominant-pole approximation is also already accurate at $n = 5$ in these cases. In Figure 4.1, we only show the result for $M = 5$, because the difference in accuracy between $M = 5$, $M = 10$ and $M = 15$ is (in these cases) almost invisible.

As can be seen from Figure 4.2, the approximations are not necessarily accurate for parameter settings in cases 3-5. Moreover, we see that large values of M result in approximations that are not close to the simulation results. In contrast to this, the estimation of the probabilities with smallest index ($n = 0, 1$) remains stable (it even improves in Figure 4.2 (c) (left)). We remark that, for visibility reasons, we have shown only the approximation for $M = 5$ in the right subplot of Figure 4.2 (c).



(a) Case 1: $\alpha = 0.587462$, $\lambda_1 = 0.355125$ and $\lambda_2 = 0.125898$.



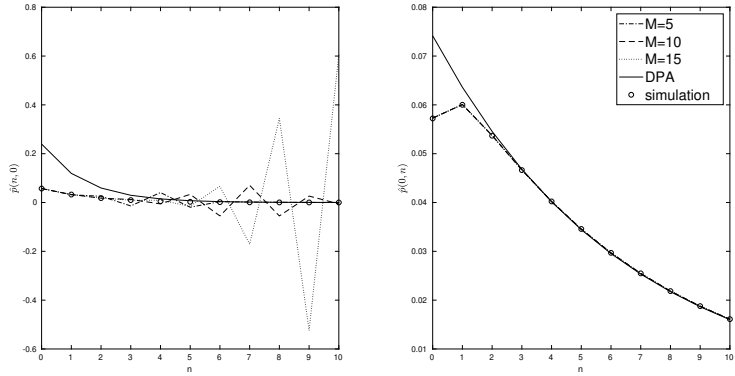
(b) Case 2: $\alpha = 0.478253$, $\lambda_1 = 0.258507$ and $\lambda_2 = 0.396966$.

Figure 4.1: Approximated probabilities $\hat{p}(n,0)$ (left) and $\hat{p}(0,n)$ (right) for truncation parameters $M = 5$.

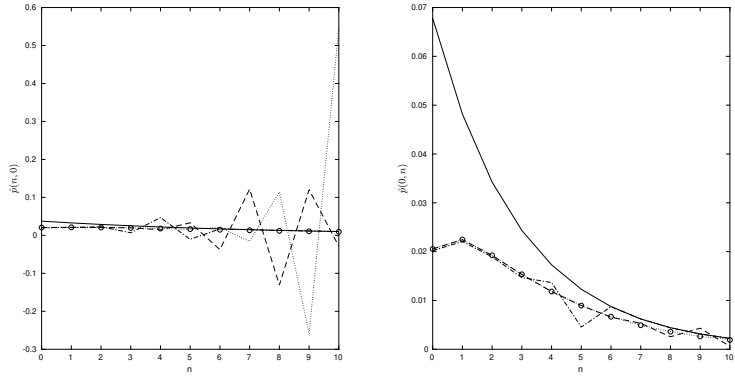
We will use these three cases to rediscuss the estimation errors of the approximation method. The bounds we can provide for the error estimation of the non-tail probabilities are too conservative (even in Case 1 and 2). Therefore, we will give an intuitive explanation of the errors at hand. We will explain this using Case 3, i.e. Figure 4.2 (a).

From the left figure in Figure 4.2 (a), it is clear that in this case the problem lies with the estimation of the coefficients $p(n,0)$ rather than $p(0,n)$. A clear difference between the two figures, is that the convergence of the dominant-pole approximation is faster for $p(0,n)$ (right) than for $p(n,0)$ (left).[1] The absolute error between this dominant-pole approximation and the simulated result suggests that the next dominant singularity of $U(z,0)$ is very close to τ_1. Hence, the factor τ_1/r_1 appearing in bound (4.9) will be very close to 1

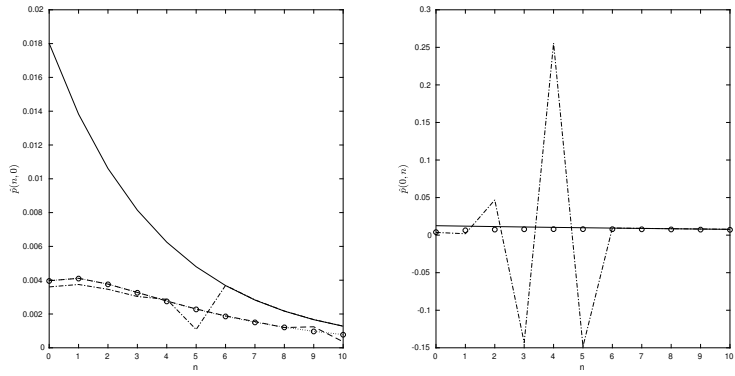
[1]Note that the scale in the two figures of Figure 4.2 (a) is different, so that all approximations are visible in the left figure.



(a) Case 3: $\alpha = 0.796212$, $\lambda_1 = 0.396695$ and $\lambda_2 = 0.177450$.

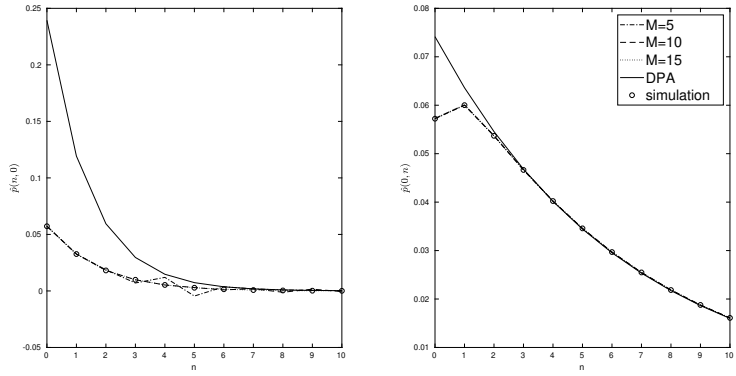


(b) Case 4: $\alpha = 0.617399$, $\lambda_1 = 0.541399$ and $\lambda_2 = 0.288171$.

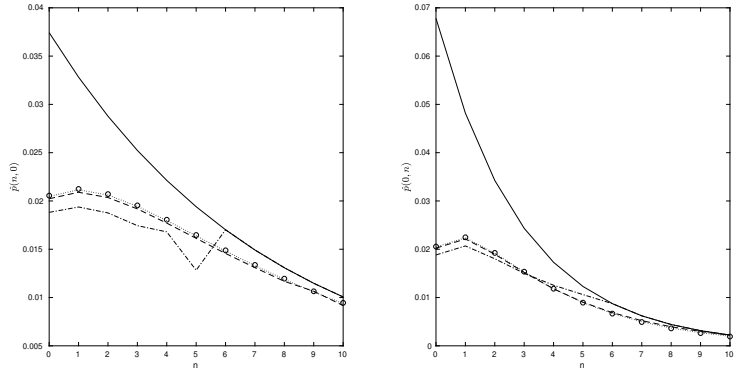


(c) Case 5: $\alpha = 0.214682$, $\lambda_1 = 0.164818$ and $\lambda_2 = 0.762360$.

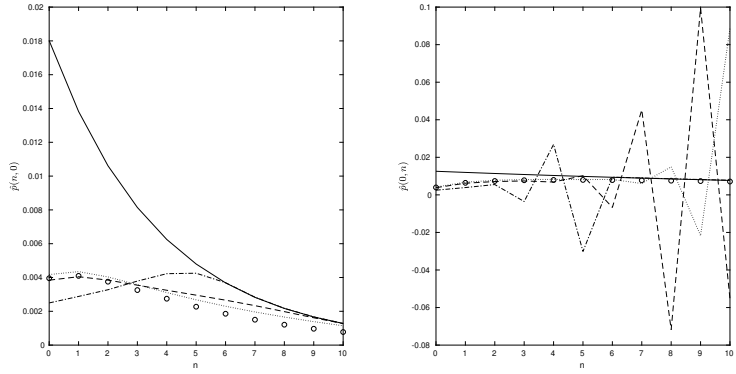
Figure 4.2: Approximated probabilities $\hat{p}(n, 0)$ (left) and $\hat{p}(0, n)$ (right) for truncation parameters $M = 5$, $M = 10$ and $M = 15$.



(a) Case 3: $\alpha = 0.796212$, $\lambda_1 = 0.396695$ and $\lambda_2 = 0.177450$.



(b) Case 4: $\alpha = 0.617399$, $\lambda_1 = 0.541399$ and $\lambda_2 = 0.288171$.



(c) Case 5: $\alpha = 0.214682$, $\lambda_1 = 0.164818$ and $\lambda_2 = 0.762360$.

Figure 4.3: Approximated probabilities $\hat{p}(n,0)$ (left) and $\hat{p}(0,n)$ (right) for truncation parameters $M = 5$, $M = 10$ and $M = 15$. The values φ_j are equidistantly chosen in $[\pi/4, \pi/2[$.

and therefore the bound will only be acceptably small for considerable large M. But, as already discussed in Sect. 4.2.2 this yields an ill-conditioned matrix T. This largely explains why the approximations for $M = 10$ and $M = 15$ are unreliable in this figure. Experiments show that the approximations for $p(0, n)$ also deteriorate from a certain M onwards, in Case 3.

Other experiments with different parameter settings confirm the behavior we just described. There are two possible solutions to overcome this problem in the future. The first possible solution is the determination of the second dominant singularity of $U(z, 0)$ and/or $U(0, z)$. A second possible solution is a different choice of φ_j, leading to different choices of zeros of the kernel. To investigate the influence of the choice of zeros of the kernel to compose our set of equations (4.23), we show in Figure 4.3 the result of the approximation method with φ_j equidistantly chosen in $[\pi/4, \pi/2[$, instead of equidistant values in $]0, \pi/2[$ as before. It is fair to say that Figure 4.3 (a) and (b) show better results when compared to Figure 4.2. This gives a minor indication that the approximation can be improved by considering other sets of zeros of the kernel. Several other choices for φ_j were also numerically investigated. Ideally, the values of φ_j are to be chosen such that the condition number of T is minimized. Because we do not have any explicit information of T^{-1} at the moment, this optimization problem is unfortunately computationally demanding and is therefore considered to be outside the scope of the present study.

Based on numerical experiments, the numerical procedure presented in this chapter is mainly applicable for light to moderate traffic regimes, and this at a very small computational cost. Regimes with a higher load cannot be tackled (with sufficient accuracy) with our approach. If one is only interested in the first few probabilities though (like $p(0, 0)$), the numerical examples show that the method is well suited for all traffic regimes.

4.4 Further discussion

In this chapter we developed a novel, original approach to approximate $U(z, 0)$ and $U(0, z)$. We used tail asymptotic results to obtain approximations for the coefficients in the Taylor series of $U(z, 0)$ and $U(0, z)$, except for the first $M + 1$ coefficients. The latter are estimated via the solution of a system of linear equations, obtained by using zero-tuples of the kernel K. However, we have seen that the approximation is not applicable to heavily loaded systems. This is mainly because too many coefficients are to be estimated in these cases. The idea of our approach is similar to the approach in [122], where a simple algorithm for the computation of the stationary distribution of a Markov chain on a semi-infinite strip is discussed. The approach in [122] exploits the geometric tail behavior of the joint queue length distribution to reduce the infinite system of equilibrium equations to a finite system of linear equations. Such an approach is easier and more satisfactory than a brute-force truncation of

the state space. A truncation approach often leads to a very large system of linear equations. In contrast to [122], we prefer to work within the framework of generating functions instead of the balance equations, since this allows us to make use of the first key step in the boundary-value technique.

Postscriptum

This chapter captures most of the results as in [98]. We mentioned in the previous paragraph that the method could be improved by determining the second dominant singularity of $U(z,0)$ and/or $U(0,z)$. This is precisely the subject of Section 3.5.2, a result we did not have when we wrote article [98]. Therefore, we discuss the influence of this new result to the approximation scheme in the paragraph below. Of particular relevance to our approach method is the approximation method recently proposed by Timmerman in [123, Chap. 6]. This method has many similarities with the one proposed in this chapter. In [123, Chap. 6], a better choice of zeros has been found. Therefore we also devote an additional paragraph to the approach as per [123, Chap. 6].

Second dominant singularity In Section 3.5.2, we obtained conditions for which it is possible to compute the second dominant singularity of $U(z,0)$ and $U(0,z)$ in case of high loads. The results of Section 3.5.2 are applicable to Case 3, Case 4 and Case 5 in this chapter. We examine Case 5 in more detail. The second dominant singularity of $U(z,0)$ can be calculated as described in Section 3.5.2. We denote with ω_1 this singularity. Then ω_1 is a simple pole of $U(z,0)$ and ω_1 is the unique solution of the following equation, cf. Theorem 3.11,

$$\omega_1\tau_2 = ((1-\alpha)\omega_1 + \alpha\tau_2)A_1(\omega_1)A_2(\tau_2), \quad \tau_1 < \omega_1 < \mathcal{R}_1 \ .$$

In Case 5, it can be verified that

$$\omega_1 = 1.352929 \ .$$

The two-pole approximation of $p(i,0)$ then reads

$$p(i,0) \sim B_0\tau_1^{-(n+1)} - F_0\omega_1^{-(n+1)} \ ,$$

with F_0 given by (3.73). We demonstrate both the dominant-pole approximation (4.5) and the two-pole approximation in Figure 4.4. From this figure, we clearly see the improvement in accuracy of the two-pole approximation in comparison with the dominant-pole approximation. It is clear that the latter converges much more slowly to the (simulated) values of $p(i,0)$.

The method developed in this chapter to approximate $U(z,0)$ could be extended as follows. Approximate $U(z,0)$ as, cf. (4.17),

$$U(z,0) \approx \hat{p}(0,0) + \sum_{n=1}^{M} \hat{p}(n,0)z^n + \frac{B_0}{\tau_1^{M+1}}\frac{z^{M+1}}{\tau_1 - z} - \frac{F_0}{\omega_1^{M+1}}\frac{z^{M+1}}{\omega_1 - z} \ .$$

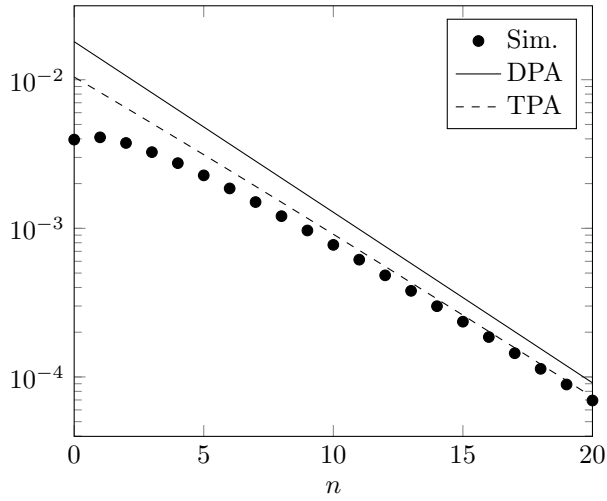


Figure 4.4: The dominant-pole approximation (DPA) and the two-pole approximation (TPA) for the probabilities $p(n, 0)$ in Case 5.

The remainder of the approximation scheme is the same as per Section 4.2.1. Broadly speaking, we now replace $U(z, 0)$ by the approximation above. The only difference in the approximation scheme is that the values s_j (4.22) change. Unfortunately, based on numerical experiments, we have observed that this extension does not yield better results for Case 3, Case 4 and Case 5.

The choice of zero-tuples as per [123] The choice of zeros seems to cause the problem in the poor approximations of Case 3, Case 4 and Case 5. The relationship between the choice of zeros and the accuracy of the approximation is also noticed recently in the PhD thesis of Timmerman [123, Ch. 6]. In this PhD thesis, a similar approximation method is developed. This method does not need the information of the tail probabilities. Broadly speaking, if we set the tail probabilities to zero in our approximation method, we obtain the method as per [123]. However, an important difference between our approaches is the choice of zero-tuples. Furthermore, it should be noted that the method in [123] is developed for a large set of multidimensional queueing models. The method is applied to several two-dimensional and multi-dimensional queueing models in Chapter 6 of [123]. In Section 6.4 of [123], Case 5 of this chapter is studied. Four different approximations, including ours, are compared. The results obtained in [123] are considerably more accurate. At first, we found this surprising. Our method is much the same as in [123], with the added exception that we make use of the tail asymptotic results wich should improve the approximation. The reason why the results in [123] are considerably better, is precisely because of the choice of zero-tuples. The results obtained in [123] are higly useful for our research. If we opt for the zero-tuples as proposed

in [123], our approximation method as we proposed in this chapter greatly improves. Therefore we will explain the choice of zeros as per [123] in some detail below.

Consider again the kernel function $K(x, y)$. First a grid structure for the y-variable is chosen. This means that a step size δ is chosen such that

$$y = \frac{k}{\delta} + \mathrm{i}\frac{l}{\delta}, \quad \text{for } k = -\delta, \ldots, -1, 1, \ldots, \delta \text{ and } l = -\delta, \ldots, -1 \ ,$$

with i the imaginary unit as before. Remark that $k, l \neq 0$. According to [123], this prevents numerical problems. Moreover l does only run until $l = -1$ (and not until δ). This is to exploit the fact that if $K(x, y) = 0$ then also $K(\bar{x}, \bar{y}) = 0$. Additionally to the grid structure for the y-variable, it is checked whether the obtained y lies within the unit circle. Next for each y an accompanying x, say $x(y)$, is computed by means of a root-finding algorithm. Finally, it is checked whether the obtained x lies within the unit circle. If so, the pair $(x(y), y)$ is added to a list of suitable zero-tuples. A while loop among increasing values of δ is constructed to ensure that sufficiently many zero-tuples are found.

The rationale behind the choice of zero-tuples in [123] is outlined in Section 6.2.3 of [123]. We echo some of the most important intuitions gathered in Section 6.2.3 of [123]. First and foremost it is observed that zero-tuples close to $(0, 0)$ are preferable. This intuition is based on the analogy with approximating the Taylor series of a function, and is confirmed experimentally. Secondly, also based on the approximation theory of one-dimensional Taylor series, it is preferred that there are x-values with positive and negative real parts. In case all zero-tuples satisfy a property like $\mathrm{Re}[x] < 0$, the author suggests to use a combination of the algorithm described in the previous paragraph and the same algorithm but with x and y interchanged.

To illustrate the gain in accuracy by choosing zeros as per [123], we compare it with our approximation results for Case 5. More concretely, we compare the following two approximations: the one that was used to obtain Figure 4.3 (c) (with $M = 15$); and our approximation method (with $M = 15$) but with zero-tuples obtained with the approach of [123]. We note that for the approach as per [123], we used the combination approach as described in the previous section. We obtain $\delta = 3$ and 32 zero pairs (16 when a grid-structure for the y-variable is chosen, and 16 when a grid-structure for the x-variable is chosen). We used the first 30 zero pairs along with the normalization equation (4.2.2). We show the two approximations in Table 4.2. From Table 4.2, we observe that more accurate approximations are obtained with this new choice of zeros. This shows yet again that the quality of the approximation strongly depends on the choice of zeros.

High-precision arithmetic Besides the choice for the zero pairs of the kernel, we have experienced that the number of digits used when making calculations may have a significant influence on the accuracy of the approximation as

	Fig. 4.3 (c)	As per [123]	Sim.
$p(0,0)$	0.004174	0.003927	0.003954
$p(1,0)$	0.004348	0.004063	0.004096
$p(2,0)$	0.004032	0.003713	0.003753
$p(3,0)$	0.003575	0.003210	0.003256
$p(4,0)$	0.003109	0.002688	0.002745
$p(5,0)$	0.002680	0.002197	0.002272
$p(6,0)$	0.002298	0.001766	0.001855
$p(7,0)$	0.001961	0.001402	0.001502
$p(8,0)$	0.001660	0.001108	0.001207
$p(9,0)$	0.001389	0.000855	0.000967
$p(10,0)$	0.001143	0.000667	0.000774
$p(0,1)$	0.004174	0.006372	0.006412
$p(0,2)$	0.006767	0.007339	0.007390
$p(0,3)$	0.007770	0.007742	0.007788
$p(0,4)$	0.008148	0.007910	0.007948
$p(0,5)$	0.008251	0.007937	0.007962
$p(0,6)$	0.008138	0.007876	0.007887
$p(0,7)$	0.008223	0.007758	0.007745
$p(0,8)$	0.006041	0.007603	0.007562
$p(0,9)$	0.014952	0.007411	0.007347
$p(0,10)$	-0.021735	0.007208	0.007113

Table 4.2: Two approximation methods for Case 5. Each row corresponds to a probability that has to be estimated. Each column corresponds to an approximation method. In the column 'Fig. 4.3 (c)' we show our approximation as in Figure 4.3 (c) with $M = 15$. In the column 'as per [123]', we show our approximation but with the suggested zeros as in [123]. Finally in the column 'Sim.', we show the simulation results.

well (with our implementation). This issue is experienced in the case of large linear systems (which corresponds with high values of M). Typically, using 50 digit precision instead of double-precision arithmetic can give rise to different (and better) results. Using high-precision arithmetic comes at the price of a higher computation time. The need of requiring high-precision arithmetic is definitely a point of concern.

Intuitively, this problem is due to the fact that the matrix T of the linear system $T\hat{\mathbf{p}} = \mathbf{s}$ is ill-conditioned. An improvement of the implementation might possibly solve this problem. A further investigation on the practical implementation is, however, beyond the scope of this chapter.

4.5 Concluding remarks

In this chapter we have proposed a novel approximation method to approximate the joint pmf of the numbers of customers in the two queues of the system. We want to remark that it is not essential for our approach to determine the marginal PGFs $U_1(z)$ and $U_2(z)$. Therefore, the applicability of our method goes beyond the queueing model studied in this dissertation. An approximate performance analysis of two-dimensional queueing models governed by Equation (2.1), for which asymptotics of $U(z_1, 0)$ and $U(0, z_2)$ are available, can be done using our approach. Such asymptotics exist for a fairly general two-class queueing model with one-step displacements only to neighboring states [86]. We want to emphasize that in this chapter, one-step displacements to non-neighboring states are allowed as well (because the two queues have batch arrivals). Even in cases where the tail distribution is not purely geometric, the method can still be applied if the corresponding residue can be found. The reason is that one only has to substitute this tail into the power series expansions.

5

Heavy-traffic analysis: a comparison study

In this chapter, we establish a heavy-traffic approximation for the correlation coefficient between the numbers of type-1 and type-2 customers in the system. The novel approximation method of Chapter 4 requires some numerical work. Moreover, when the load is high, the approximation scheme leads to an ill-conditioned set of linear equations. Besides the numerical motive, we are also interested in understanding more about the behavior of the system in the case of heavy traffic.

The heavy-traffic result is computed via the solution of a two-dimensional functional equation, obtained by formulating a boundary-value problem on a hyperbola. Additionally, we compare our model with a related model. In most server-sharing models, it is assumed that the system is work-conserving in the sense that if one of the queues is empty, a customer of the other queue is served with probability 1. In the second model, say the modified model, we include this work-conserving rule such that the server is always allocated to a non-empty queue. Contrarily to what we would expect, the resulting heavy-traffic approximations reveal that both models remain different for critically loaded queues.

As already elaborated in Section 1.5, the marginal distributions of the numbers of customers present in both queues are easy to calculate for the original model. In contrast to this, in the case of the modified model, the distribution of the *total* number of customers is easy to calculate since the total system is identical to a single-server single-queue model. These two observations allow for comparing both models by means of a simple mean value analysis. We are however also interested in joint performance measures, in particular the correlation coefficient between the two system contents. Obtaining joint performance measures brings us back to the machinery of the boundary-value method, of which we already argued in Chapter 2 that it rarely provides explicit, let alone easy to calculate, performance measures.

In this chapter, we will make the simplifying assumption of symmetry between the two queues. In our case, the heavy-traffic limit means that we will let the mean arrival rates go to their respective critical value of instability such that the queues are nearly saturated. The reason for considering the heavy-traffic limit is twofold. The first reason is that by considering the heavy-traffic limit, the boundary-value method can be applied in a simpler manner compared to the non-heavy-traffic case, such that the numerical work is limited (modified model) or even not necessary (original model). Moreover, since it is typical for heavy-traffic results to be rather insensitive to the exact form of the arrival (and service) process [124], we can assume a general batch arrival process for our queueing models. The second reason is that in heavy-traffic, we have the interesting question whether (or not) both models converge to the same model. It is reasonable to think that the answer to this question is affirmative since it is expected that the queues empty less and less near saturation, and hence the only rule that sets the two models apart applies less frequently. However, we show in this chapter that this is not true.

Earlier studies that consider the heavy-traffic limit of similar, continuous-time, queueing models are [125–127]. It is worth mentioning that these papers use a heavy-traffic diffusion approximation (see for example also [128] and [129] for a more detailed explanation of this method). Broadly speaking, with the diffusion approximation one replaces the balance equations by a diffusion equation with appropriate boundary conditions. Recently, in the doctoral thesis [130, Ch. 4], a heavy-traffic limit is obtained using transforms, similar as in the classic work of [124], but with the added difficulty that still a boundary-value problem for analytic functions has to be solved. However, a *nice* boundary-value problem is obtained, where the boundary consist of a parabola, for which an explicit solution is obtained. We note that the model in [130, Ch. 4] is a (continuous-time) two-queue random time-limited Markov modulated fluid polling model, and is in a sense somewhat similar to the model as described in Section 1.4. Recently in [131], the transient process-limit of the joint workload in heavy traffic is investigated under less restrictive assumptions on the input process (no longer constant fluid flows) and the server switching process (no longer assumed to be exponential). By solving a boundary-value problem, the stationary distribution of the limiting process is determined.

This chapter follows the lines of our contribution [132]. In Section 5.1, we describe the assumptions, notations and definitions used in this chapter. Section 5.2 presents the problem statement and lists the main results of this chapter. In Section 5.3, the analysis for our model is carried out. Finally, Section 5.4 is devoted to the analysis of the work-conserving variant of our model. In both Section 5.3 and Section 5.4, we also discuss numerical results.

5.1 Mathematical model and preliminary results

In this section, we introduce the assumptions, notations and definitions used in the remainder of the chapter.

We consider a discrete-time single-server queueing system with two infinitely sized queues, say queue-1 and queue-2, and two independent input lines. Customers arriving at queue-1 and queue-2 are referred to as type-1 and type-2 customers respectively. Time is assumed to be slotted. The numbers of type-1 and type-2 arrivals during slot k are denoted by $a_{1,k}$ and $a_{2,k}$ respectively. The sequences of discrete random variables $(a_{1,k})_{k\in\mathbb{N}}$ and $(a_{2,k})_{k\in\mathbb{N}}$ are assumed to be i.i.d. and are specified by a common probability generating function (pgf) $A(z)$, i.e.

$$A(z) \triangleq \mathsf{E}[z^{a_{j,k}}], \quad j = 1, 2 \, .$$

We thus assume that both types of customers have the same arrival process, i.e. $a_{1,k}$ and $a_{2,k}$ have the same probability distribution. Moreover, we assume that these two arrival processes are independent, i.e. the random variables $a_{1,k}$ and $a_{2,k}$ are independent. For further use, let λ denote the mean numbers of type-1 and type-2 arrivals per slot,

$$\lambda \triangleq A'(1) \, . \tag{5.1}$$

At the beginning of every slot, the single-server selects either queue with probability $\frac{1}{2}$. In case a non-empty queue is selected, a customer of this queue gets served. We assume that the service of each customer type requires exactly one slot, regardless of whether the customer is of type 1 or type 2. Based on what happens when an empty queue is selected, we make the distinction between the following two service disciplines.

Service discipline 1 (Non-work-conserving policy)**.** *We assume that when an empty queue is chosen, no service occurs in that slot, even when the other queue is non-empty. We will refer to this service discipline as the non-work-conserving policy.*

Service discipline 2 (Work-conserving policy)**.** *When only one of both queues is non-empty and the other is empty, the non-empty queue is served during that slot, even when initially the empty queue was selected. We will refer to this service discipline as the work-conserving policy.*

Remark that the first service discipline is the same as described in Section 1.4, but with $\alpha = \frac{1}{2}$.

We now write down the key functional equations of the joint PGF of the system contents.

5.1.1 The non-work-conserving policy

Let u_1 and u_2 indicate the number of type-1 and type-2 customers respectively, in steady-state. We define their joint PGF as

$$U(z_1, z_2) \triangleq \mathsf{E}[z_1^{u_1} z_2^{u_2}] \ . \tag{5.2}$$

Because of the non-work-conserving property, the queues can be analyzed separately, cf. Section 1.5. Both queues are then equivalent to simple single-server queues with geometric service times with mean 2. The arrival load offered by type-j customers is given by 2λ. The stability condition of a single queue is given by

$$2\lambda < 1 \ , \tag{5.3}$$

which is therefore also the stability condition of the complete queueing system.

Recall that we found a functional equation for the joint PGF of the number of type-1 and type-2 customers in Chapter 2, equation (2.12), for the general setting with different arrival distributions for type-1 and type-2 customers and the probability that the server selects queue-1 is parameterized by α. Accounting for the symmetry in this chapter, the functional equation for this chapter reads:

$$K(z_1, z_2)U(z_1, z_2) = A(z_1)A(z_2)\frac{1}{2}[(z_2 - 1)z_1 U(z_1, 0) + (z_1 - 1)z_2 U(0, z_2)] \ , \tag{5.4}$$

with

$$K(z_1, z_2) = z_1 z_2 - \frac{1}{2}[z_1 + z_2]A(z_1)A(z_2) \ . \tag{5.5}$$

As usual, we will refer to $K(z_1, z_2)$ as the kernel of equation (5.4).

Equation (5.4) is a functional equation for $U(z_1, z_2)$ as the boundary functions $U(z_1, 0)$ and $U(0, z_2)$ are present in the RHS of the equation. These boundary functions can be found by constructing a boundary-value problem for analytic functions. As already mentioned in Section 2.1, this will require a numerical approach. In this chapter, we however pursue a different objective: we propose a heavy-traffic approximation that requires only a minimum amount of (or no) numerical computations. We postpone this analysis to Section 3 and 4. First we give some, more elementary, results of the joint PGF $U(z_1, z_2)$.

It is not difficult to see that the joint PGF $U(z_1, z_2)$ of the system contents is symmetric in z_1 and z_2, i.e.

$$U(z_1, z_2) = U(z_2, z_1) \ .$$

In particular, this implies that there is only one boundary function present in the functional equation for $U(z_1, z_2)$, since

$$U(z, 0) = U(0, z) \ . \tag{5.6}$$

The PGF $S(z)$ of the total number of customers in the system can be expressed in terms of this boundary function. Substituting $z_1 = z_2 = z$ into (5.4) yields

$$S(z) \triangleq U(z,z) = \frac{A^2(z)(z-1)U(z,0)}{z - A^2(z)} \,. \tag{5.7}$$

On the other hand, the marginal PGFs $U(z,1)$ and $U(1,z)$ of the numbers of type-1 and type-2 customers do not depend on $U(z,0)$. Moreover $U(z,1)$ and $U(1,z)$ are identical and given by

$$U(z,1) = \frac{A(z)(z-1)(\frac{1}{2} - \lambda)}{z - \frac{1}{2}(z+1)A(z)} \,. \tag{5.8}$$

A first important performance measure that can be derived from this PGF, is the probability of an empty queue at the beginning of a random slot:

$$\mathsf{P}[u_1 = 0] = \mathsf{P}[u_2 = 0] = U(0,1) = 1 - 2\lambda \,. \tag{5.9}$$

Further, all the marginal moments of interest of the random variables u_1 and u_2 can thus be computed. For example, the mean is given by

$$\begin{aligned} \mathsf{E}[u_1] = \mathsf{E}[u_2] &= \left. \frac{dU(z,1)}{dz} \right|_{z=1} \\ &= \lambda + \frac{\lambda + A''(1)}{1 - 2\lambda} \,. \end{aligned} \tag{5.10}$$

5.1.2 The work-conserving policy

Let v_1 and v_2 be the number of type-1 and type-2 customers in steady-state. Their joint PGF is defined by

$$V(z_1, z_2) \triangleq \mathsf{E}[z_1^{v_1} z_2^{v_2}] \,. \tag{5.11}$$

This model is a special case of the model studied in [18]. Introducing the assumptions of the present paper (i.e. single slot service times, the same arrival distributions and weights), the following functional equation for $V(z_1, z_2)$ is obtained:

$$\begin{aligned} &K(z_1,z_2)V(z_1,z_2) = \\ &\quad A(z_1)A(z_2)\frac{1}{2}\left((z_2 - z_1)(V(z_1,0) - V(0,z_2)) + V(0,0)(2z_1z_2 - z_1 - z_2)\right) \,, \end{aligned} \tag{5.12}$$

where K is given by (5.5). Note that both functional equations have the same kernel.

In case of a work-conserving service discipline and single-slot service times, the distribution of the total number of customers is easy to calculate since the total system is identical to a single-server model with a single queue. Indeed, set $z_1 = z_2 = z$ in (5.12) to obtain

$$V(z,z) = \frac{A^2(z)V(0,0)(z-1)}{z - A^2(z)} , \tag{5.13}$$

and then take the limit $z \to 1$ to obtain (using l'Hôpital's rule)

$$V(0,0) = 1 - 2\lambda . \tag{5.14}$$

All the moments of the total system content $v_1 + v_2$ can therefore be computed. For example, we obtain that the mean total number of customers in the system is given by

$$\begin{aligned} \mathsf{E}[v_1 + v_2] &= \left. \frac{dV(z,z)}{dz} \right|_{z=1} \\ &= 2\lambda + \frac{\lambda^2 + A''(1)}{1 - 2\lambda} . \end{aligned} \tag{5.15}$$

It is not difficult to see that the joint pgf $V(z_1, z_2)$ is symmetric in z_1, z_2, i.e.

$$V(z_1, z_2) = V(z_2, z_1) .$$

As a consequence, there is only one boundary function present in equation (5.12). Moreover, the marginal pgfs $V(z,1)$ and $V(1,z)$ of the system contents of type-1 and type-2 customers are identical and given by

$$Q(z) \triangleq V(z,1) = V(1,z) = \frac{A(z)(z-1)(1 - 2\lambda + V(0,1) - V(z,0))}{2z - (z+1)A(z)} , \tag{5.16}$$

where we used (5.14). It is worth noting that (5.14) indicates the probability that the total system is empty at the beginning of a random slot. Because of the work-conserving service discipline, the stability condition of the system is naturally given by $2\lambda < 1$, i.e. (5.3). This is the same stability condition as for the non-work-conserving policy. This is not a coincidence, and it is precisely therefore that we assumed a symmetrical system. We emphasize that in the case of a non-symmetrical system, both systems have a different stability condition (cf. Section 1.5 and [48, Sect. 4.2]).

5.2 Problem statement and main results

We are interested in the influence of the service discipline on the correlation structure between the numbers of type-1 and type-2 customers, when λ is near

the critical value $\frac{1}{2}$. The simplest possible performance measure to quantify the correlation structure is the correlation coefficient between the number of type-1 and type-2 customers. The objective of our analysis is to obtain expressions for $\lim_{\lambda\uparrow\frac{1}{2}} \text{corr}[u_1, u_2]$ and $\lim_{\lambda\uparrow\frac{1}{2}} \text{corr}[v_1, v_2]$. Our approach is to obtain the joint Laplace-Stieltjes transform (LST) of the scaled system contents $(1-2\lambda)u_1$ and $(1-2\lambda)u_2$ as $\lambda \uparrow \frac{1}{2}$ and the joint LST of the scaled system contents $(1-2\lambda)v_1$ and $(1-2\lambda)v_2$ as $\lambda \uparrow \frac{1}{2}$. Notice that for two random variables X and Y

$$\text{corr}[X, Y] = \text{corr}[(1-2\lambda)X, (1-2\lambda)Y] \,. \tag{5.17}$$

The importance of equation (5.17) lies in the fact that if we are able to compute the RHS for $\lambda \to \frac{1}{2}$, then we have also obtained the LHS for $\lambda \to \frac{1}{2}$.

Functional equations for the two limiting joint LSTs can easily be obtained from (5.4) and (5.12). The kernel in these new functional equations is much simpler than $K(z_1, z_2)$, as will be shown in Section 5.3.

In this chapter, we assume that $A''(1)$ has a finite limit as $\lambda \uparrow \frac{1}{2}$. We define

$$\lambda_{11} = \lim_{\lambda\uparrow\frac{1}{2}} A''(1) \,. \tag{5.18}$$

This mathematical quantity can be expressed in terms of the physical quantity $\text{var}[a_{j,k}]$, since $\text{var}[a_{j,k}] = A''(1) + \lambda - \lambda^2$. Therefore, we define σ_h^2 as the variance of the number of type-j arrivals per slot, as $\lambda \uparrow \frac{1}{2}$, i.e.

$$\sigma_h^2 \triangleq \lim_{\lambda\uparrow\frac{1}{2}} \text{var}[a_{j,k}] = \lambda_{11} + \frac{1}{4} \,. \tag{5.19}$$

Notice that σ_h^2 is the same for both customer types, because of the symmetry of the arrival process. Since $A(z)$ is a power series in z with positive coefficients, we have that $\lambda_{11} \geq 0$. Consequently, we have the following important lower bound for σ_h^2 :

$$\sigma_h^2 \geq \frac{1}{4} \,. \tag{5.20}$$

When λ approaches its critical value $\frac{1}{2}$ in the non-work-conserving model, it is expected that the probability to select an empty server tends to zero because $\mathsf{P}[u_1 = 0] = \mathsf{P}[u_2 = 0] = 1 - 2\lambda \downarrow 0$, as $\lambda \uparrow \frac{1}{2}$. In that perspective, the non-work-conserving model resembles the work-conserving model when λ is close to $\frac{1}{2}$, because the number of wasted slots in the non-work-conserving model decreases. However, a simple mean value analysis already reveals that even in the limit $\lambda \uparrow \frac{1}{2}$ the work-conserving model is still strictly more efficient. Indeed, from (5.10) we have that when $\lambda \uparrow \frac{1}{2}$ the mean total scaled system content $(1-2\lambda)(u_1 + u_2)$ tends to

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}\left[(1-2\lambda)\,(u_1+u_2)\right] = 2\left(\frac{1}{2} + \lambda_{11}\right)$$

$$= \frac{1}{2} + 2\sigma_h^2 \,, \tag{5.21}$$

while from (5.15), we easily obtain that

$$\begin{aligned} \lim_{\lambda \uparrow \frac{1}{2}} \mathsf{E}\left[(1-2\lambda)\,(v_1+v_2)\right] &= \frac{1}{4} + \lambda_{11} \\ &= \sigma_h^2 \,. \end{aligned} \tag{5.22}$$

Hence, we have that for $\lambda \uparrow \frac{1}{2}$

$$\mathsf{E}[u_1+u_2] \sim \frac{\frac{1}{2}+2\sigma_h^2}{1-2\lambda} \quad \text{and} \quad \mathsf{E}[v_1+v_2] \sim \frac{\sigma_h^2}{1-2\lambda} \,, \tag{5.23}$$

where $f(x) \sim g(x)$ means that $\frac{f(x)}{g(x)}$ tends to 1 as $x \to x_0$. Consequently, we have shown that the mean total system contents are asymptotically not equivalent.

In Sections 5.3 and 5.4 we obtain the limiting LSTs of the scaled system contents. From these expressions, the correlation coefficient for $\lambda \uparrow \frac{1}{2}$ under the different service disciplines is obtained. The final, complete expressions are given in the theorem below.

Theorem 5.1. *The correlation coefficient between the numbers of type-1 and type-2 customers in the system with the non-work-conserving policy, for $\lambda \uparrow \frac{1}{2}$ is given by*

$$\lim_{\lambda \uparrow \frac{1}{2}} \operatorname{corr}[u_1, u_2] = \frac{32\pi^2\sigma_h^4 - 128\sigma_h^4\varphi^2 + 16\pi^2\sigma_h^2 - 112\sigma_h^2\varphi^2 - 21\varphi^2}{3\varphi^2\left(1+4\sigma_h^2\right)^2} \,,$$

where

$$\varphi = \arccos\left(\frac{1}{1+4\sigma_h^2}\right) \,. \tag{5.24}$$

The correlation coefficient between the numbers of type-1 and type-2 customers in the system with the work-conserving policy, for $\lambda \uparrow \frac{1}{2}$ is given by

$$\lim_{\lambda \uparrow \frac{1}{2}} \operatorname{corr}[v_1, v_2] = \frac{\sigma_h^4}{2}\left(\frac{3}{4}\sigma_h^4 + \frac{1}{4}\sigma_h^2 + J\right)^{-1} - 1 \,, \tag{5.25}$$

where

$$J = \int_0^{+\infty} \frac{r(t)}{(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1)^2} \cdot \frac{\tanh\left(\frac{\pi}{\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)\left(1+2\cosh(t)k\sigma_h^2\right)} dt \,, \tag{5.26}$$

$$k = \frac{1}{2\sigma_h^2}\sqrt{\frac{1+4\sigma_h^2}{1+2\sigma_h^2}} \,, \tag{5.27}$$

$$r(t) = \frac{\pi}{\varphi^2}(1+2\sigma_h^2)^{\frac{3}{2}}\sqrt{1+4\sigma_h^2}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1\right) + \frac{\pi^2}{\varphi^3}2\sqrt{2}\sigma_h(1+2\sigma_h^2)\sqrt{1+4\sigma_h^2}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1\right) . \tag{5.28}$$

The result of Theorem 5.1 is discussed in Sections 5.3.4 and 5.4.4, but we already mention that both correlation coefficients are significantly different when compared to each other.

5.3 The non work-conserving policy in heavy-traffic

The purpose of the analysis in this section is to obtain the joint LST of the scaled system contents $(1-2\lambda)u_1$ and $(1-2\lambda)u_2$ as $\lambda \uparrow \frac{1}{2}$.

The joint LST of $(1-2\lambda)u_1$ and $(1-2\lambda)u_2$ is

$$\begin{aligned} U_\lambda(s_1,s_2) &\triangleq \mathsf{E}[e^{-s_1(1-2\lambda)u_1 - s_2(1-2\lambda)u_2}] \\ &= U(e^{-s_1(1-2\lambda)}, e^{-s_2(1-2\lambda)}) . \end{aligned}$$

Hence, from (5.4) we obtain an equation for $U_\lambda(s_1,s_2)$:

$$\begin{aligned} &2K\Big(e^{-s_1(1-2\lambda)}, e^{-s_2(1-2\lambda)}\Big)U_\lambda(s_1,s_2) = \\ &e^{-s_1(1-2\lambda)}\Big(e^{-s_2(1-2\lambda)}-1\Big)A\Big(e^{-s_1(1-2\lambda)}\Big)A\Big(e^{-s_2(1-2\lambda)}\Big)U\Big(e^{-s_1(1-2\lambda)},0\Big) \\ &+ e^{-s_2(1-2\lambda)}\Big(e^{-s_1(1-2\lambda)}-1\Big)A\Big(e^{-s_1(1-2\lambda)}\Big)A\Big(e^{-s_2(1-2\lambda)}\Big)U\Big(0,e^{-s_2(1-2\lambda)}\Big) . \end{aligned} \tag{5.29}$$

We assume that the following limit exists:

$$\begin{aligned} U_h(s_1,s_2) &\triangleq \lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}[e^{-s_1(1-2\lambda)u_1 - s_2(1-2\lambda)u_2}] \\ &= \lim_{\lambda\uparrow\frac{1}{2}} U_\lambda(s_1,s_2) . \end{aligned} \tag{5.30}$$

The subscript h is to indicate that the LST corresponds to the heavy-traffic limit. We also define

$$\begin{aligned} S_h(s) &\triangleq \lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}[e^{-s(1-2\lambda)(u_1+u_2)}] \\ &= U_h(s,s) \\ &= \lim_{\lambda\uparrow\frac{1}{2}} S(e^{-s(1-2\lambda)}) . \end{aligned} \tag{5.31}$$

as the limiting LST of the total scaled system content.

The boundary function $U(e^{-s(1-2\lambda)},0)$, as $\lambda \uparrow \frac{1}{2}$, can be written in terms of $S_h(s)$. From (5.7), we obtain that

$$\frac{zU(z,0)}{z-1} = \frac{z(z-A^2(z))}{A^2(z)(z-1)^2}S(z)\ . \tag{5.32}$$

We have to use the change of variables $z = e^{-s(1-2\lambda)}$ in this expression and take the limit for $\lambda \uparrow \frac{1}{2}$. However, we have to be careful with derivatives of $A(e^{-s(1-2\lambda)})$ with respect to λ since $A(\cdot)$ itself also depends on λ. For ease of presentation, we therefore make this dependency on λ explicit and write $A(\lambda,z)$, instead of $A(z)$. Therefore, if we substitute $z = e^{-s(1-2\lambda)}$ in (5.32) and take the limit for $\lambda \uparrow \frac{1}{2}$ we obtain that

$$\begin{aligned}
&\lim_{\lambda\uparrow\frac{1}{2}} \frac{e^{-s(1-2\lambda)}U(e^{-s(1-2\lambda)},0)}{e^{-s(1-2\lambda)}-1} \\
&= \lim_{\lambda\uparrow\frac{1}{2}} \frac{e^{-s(1-2\lambda)}(e^{-s(1-2\lambda)}-A^2(\lambda,e^{-s(1-2\lambda)}))}{A^2(\lambda,e^{-s(1-2\lambda)})(e^{-s(1-2\lambda)}-1)^2}S(e^{-s(1-2\lambda)}) \\
&= S_h(s)\cdot\lim_{\lambda\uparrow\frac{1}{2}} \frac{e^{-s(1-2\lambda)}-A^2(\lambda,e^{-s(1-2\lambda)})}{(e^{-s(1-2\lambda)}-1)^2}\ .
\end{aligned} \tag{5.33}$$

Note that $A(\lambda,1)=1$ for all λ, because of the normalization condition of a PGF. Hence, $e^{-s(1-2\lambda)}-A^2(\lambda,e^{-s(1-2\lambda)})$ goes to zero as $\lambda\uparrow\frac{1}{2}$. Therefore, we will have to apply l'Hôpital's rule at least once to the limit in (5.33). As we will show further, the following partial derivatives will occur:

$$A^{(1)}(\lambda,1) = 0\ , \tag{5.34}$$

$$A^{(11)}(\lambda,1) = 0\ , \tag{5.35}$$

$$A^{(2)}(\lambda,1) = \lambda\ , \tag{5.36}$$

$$A^{(22)}(\lambda,1) = A''(1)\ , \tag{5.37}$$

$$A^{(12)}(\lambda,1) = 1\ , \tag{5.38}$$

with

$$A^{(\mathbf{1}_n,\mathbf{2}_m)}(x,y) \triangleq \frac{\partial^n}{\partial\lambda^n}\frac{\partial^m}{\partial z^m}A(\lambda,z)\ ,\Big|_{\lambda=x,z=y}\ , \tag{5.39}$$

whereby $\mathbf{k}_n$ represents a series consisting of n consecutive k's.

Perhaps the easiest way to obtain (5.34)-(5.38), is to look at a series expansion of $A(\lambda,z)$ about $z=1$. Since the first two moments of $A(z)$ exist, we can write the following series expansion of $A(\lambda,z)$ about $z=1$ for fixed λ:

$$\begin{aligned}
A(\lambda,z) &= A(\lambda,1) + A^{(2)}(\lambda,1)(z-1) + \frac{1}{2}A^{(22)}(\lambda,1)(z-1)^2 + O\left((z-1)^3\right) \\
&= 1 + \lambda(z-1) + \frac{1}{2}A^{(22)}(\lambda,1)(z-1)^2 + O\left((z-1)^3\right)\ .
\end{aligned} \tag{5.40}$$

If we differentiate (5.40) with respect to λ once or twice, all terms have a common factor $(z-1)$. Hence, by then evaluating at $z=1$, we get (5.34) and (5.35). Equations (5.36) and (5.37) follow from the first two moments of $A(z)$, where we used definition (5.1) to simplify (5.36). Finally, equation (5.38) can be found by first taking the derivative of (5.40) with respect to z, yielding $\lambda + O(z-1)$, and then taking the derivative with respect to λ. Thereafter, by evaluating at $z=1$ we obtain (5.38).

For further use, we remark that

$$\lim_{\lambda\uparrow\frac{1}{2}} A^{(\mathbf{1}_n,\mathbf{2}_m)}(\lambda, e^{-s(1-2\lambda)}) = \lim_{\lambda\uparrow\frac{1}{2}} A^{(\mathbf{1}_n,\mathbf{2}_m)}(\lambda, 1)\ , \tag{5.41}$$

since $A^{(\mathbf{1}_n,\mathbf{2}_m)}(\lambda, e^{-s(1-2\lambda)})$ can be written as a sum of products of $(e^{-s(1-2\lambda)}-1)^k$ with a factor that depends on λ, cf. (5.40).

Returning to the limit in (5.33), we have that

$$\begin{aligned}
&\lim_{\lambda\uparrow\frac{1}{2}} \frac{e^{-s(1-2\lambda)} - A^2(\lambda, e^{-s(1-2\lambda)})}{(e^{-s(1-2\lambda)}-1)^2} \\
=&\lim_{\lambda\uparrow\frac{1}{2}} \frac{1}{4s(e^{-s(1-2\lambda)}-1)e^{-s(1-2\lambda)}} \cdot \Big\{ 2se^{-s(1-2\lambda)} \\
&-2A(\lambda, e^{-s(1-2\lambda)}) \left(A^{(1)}(\lambda, e^{-s(1-2\lambda)}) + 2se^{-s(1-2\lambda)} A^{(2)}(\lambda, e^{-s(1-2\lambda)}) \right) \Big\} \\
=&\lim_{\lambda\uparrow\frac{1}{2}} \frac{1}{8s^2(e^{-s(1-2\lambda)})^2 + 8(e^{-s(1-2\lambda)}-1)s^2e^{-s(1-2\lambda)}} \\
&\cdot \Big\{ 4s^2e^{-s(1-2\lambda)} - 2\left(A^{(1)}(\lambda, e^{-s(1-2\lambda)}) + 2se^{-s(1-2\lambda)} A^{(2)}(\lambda, e^{-s(1-2\lambda)}) \right)^2 \\
&-2A(\lambda, e^{-s(1-2\lambda)}) \Big[A^{(11)}(\lambda, e^{-s(1-2\lambda)}) + 4se^{-s(1-2\lambda)} A^{(12)}(\lambda, e^{-s(1-2\lambda)}) \\
&+4s^2e^{-s(1-2\lambda)} A^{(2)}(\lambda, e^{-s(1-2\lambda)}) + (2se^{-s(1-2\lambda)})^2 A^{(22)}(\lambda, e^{-s(1-2\lambda)}) \Big] \Big\}\ . \\
=&\frac{-2s^2 - 8s - 8\lambda_{11}s^2}{8s^2} \\
=&-\left(\sigma_h^2 + \frac{1}{s} \right)\ .
\end{aligned}$$

In the second equality we used l'Hôpital's rule again, since (5.34) and (5.36) imply that the numerator in the first equality goes to zero. In the third equality, we used (5.34)-(5.38) and (5.18). The final equality follows from (5.19). Substituting this result into (5.33) gives us

$$\lim_{\lambda\uparrow\frac{1}{2}} \frac{e^{-s(1-2\lambda)} U(e^{-s(1-2\lambda)}, 0)}{e^{-s(1-2\lambda)}-1} = -\left(\sigma_h^2 + \frac{1}{s} \right) S_h(s)\ . \tag{5.42}$$

We can now proceed with the determination of a functional equation for $U_h(s_1, s_2)$. Dividing (5.29) by $\left(e^{-s_1(1-2\lambda)}-1\right)\left(e^{-s_2(1-2\lambda)}-1\right)$, taking the

limit $\lambda \uparrow \frac{1}{2}$ and using (5.42), we obtain that

$$\frac{K_h(s_1,s_2)}{4s_1s_2}U_h(s_1,s_2) = \left(\sigma_h^2 + \frac{1}{s_1}\right) S_h(s_1) + \left(\sigma_h^2 + \frac{1}{s_2}\right) S_h(s_2)\,, \qquad (5.43)$$

with

$$K_h(s_1,s_2) \triangleq (1+4\sigma_h^2)(s_1^2+s_2^2) - 2s_1s_2 + 4(s_1+s_2) \qquad (5.44)$$

as the kernel of the functional equation (5.43).

It is easily seen that the marginal LSTs $U_h(s,0)$ and $U_h(0,s)$ can be obtained by choosing either $\{s_1 = s, s_2 = 0\}$ or $\{s_1 = 0, s_2 = s\}$ in equation (5.43). We obtain that

$$U_h(s,0) = \frac{4}{(1+4\sigma_h^2)\,s + 4}, \qquad (5.45)$$

which is the LST of the distribution of an exponential random variable with mean $\sigma_h^2 + \frac{1}{4}$. Clearly, we have the same result for $U_h(0,s)$.

5.3.1 Areas of convergence

In this section, we examine in which region(s) the LSTs $U_h(s_1,s_2)$ and $S_h(s)$ are analytic. From Laplace transform theory, we have that $U_h(s,0)$ is analytic for $\mathrm{Re}[s] > -\frac{4}{1+4\sigma_h^2}$ and so we may conclude that $U_h(s_1,s_2)$ is at least analytic in the region

$$\left\{(s_1,s_2) : \mathrm{Re}[s_1] > -\frac{4}{1+4\sigma_h^2}, \mathrm{Re}[s_2] \geq 0\right\} \cup \left\{(s_1,s_2) : \mathrm{Re}[s_1] \geq 0, \mathrm{Re}[s_2] > -\frac{4}{1+4\sigma_h^2}\right\}. \qquad (5.46)$$

Since $U_h(s_1,s_2)$ is joint analytic inside this region, this also holds for $K_h(s_1,s_2)U_h(s_1,s_2)$. Finally, using functional equation (5.43) we can conclude that $S_h(s)$ must be analytic for $\mathrm{Re}[s] > -\frac{4}{1+4\sigma_h^2}$.

5.3.2 Solution of the functional equation

In this section, we solve the functional equation (5.43) for $U_h(s_1,s_2)$. In order to solve this equation, we have to determine the function $S_h(s)$. This can be done by exploiting the fact that when $K_h(s_1,s_2)$ vanishes for a certain pair (s_1,s_2) for which $U_h(s_1,s_2)$ is finite, it must be that the RHS of (5.43) vanishes for that pair (s_1,s_2).

Observe that $K_h(s_1,s_2)$ is a quadratic polynomial both in s_1 and s_2, where s_1 and s_2 may be complex-valued. For any given value s_1 there exist exactly two values, say $\tilde{s}_2$ and $\hat{s}_2$, such that $K_h(s_1,\tilde{s}_2) = 0$, $K_h(s_1,\hat{s}_2) = 0$. Using

the well-known relations for the product and sum of the roots of a quadratic equation, we always have that

$$\begin{aligned} \tilde{s}_2\hat{s}_2 &= \frac{(4+(1+4\sigma_h^2)s_1)s_1}{1+4\sigma_h^2} \,, \\ \tilde{s}_2+\hat{s}_2 &= \frac{2s_1-4}{1+4\sigma_h^2} \,, \end{aligned} \tag{5.47}$$

since

$$K_h(s_1,s_2) = (1+4\sigma_h^2)s_2^2 + (4-2s_1)s_2 + (4+(1+4\sigma_h^2)s_1)s_1 \,.$$

If $U_h(s_1,\tilde{s}_2)$ and $U_h(s_1,\hat{s}_2)$ are finite, we obtain from (5.43) that

$$\left(\sigma_h^2+\frac{1}{s_1}\right)S_h(s_1) + \left(\sigma_h^2+\frac{1}{\tilde{s}_2}\right)S_h(\tilde{s}_2) = 0 \,,$$

and

$$\left(\sigma_h^2+\frac{1}{s_1}\right)S_h(s_1) + \left(\sigma_h^2+\frac{1}{\hat{s}_2}\right)S_h(\hat{s}_2) = 0 \,,$$

so that by eliminating $S_h(s_1)$

$$\left(\sigma_h^2+\frac{1}{\tilde{s}_2}\right)S_h(\tilde{s}_2) = \left(\sigma_h^2+\frac{1}{\hat{s}_2}\right)S_h(\hat{s}_2) \,. \tag{5.48}$$

We consider zeros such that $\overline{\hat{s}_2} = \tilde{s}_2$. As we will see further, this choice of zeros is sufficient to determine $S_h(s)$. Letting

$$z = x+\mathrm{i}y \in \mathbb{C}, \quad \tilde{s}_2 = z \quad \text{and} \quad \hat{s}_2 = \bar{z} \,, \tag{5.49}$$

we will first prove the existence and the exact location of such zeros. Subsituting $\tilde{s}_2 = x+\mathrm{i}y$ and $\hat{s}_2 = x-\mathrm{i}y$ into (5.47) yields

$$x^2+y^2 = \frac{(4+(1+4\sigma_h^2)s_1)s_1}{1+4\sigma_h^2} \,, \tag{5.50}$$

and

$$2x = \frac{2s_1-4}{1+4\sigma_h^2} \,. \tag{5.51}$$

Solving (5.51) for s_1 yields

$$s_1 = (1+4\sigma_h^2)x+2 \,, \tag{5.52}$$

and substituting this into (5.50) gives us

$$x^2+y^2 = \frac{\left(4\,\sigma_h^2+1\right)\left(4\,\sigma_h^2x+x+2\right)^2+16\,\sigma_h^2x+4\,x+8}{4\,\sigma_h^2+1} \,.$$

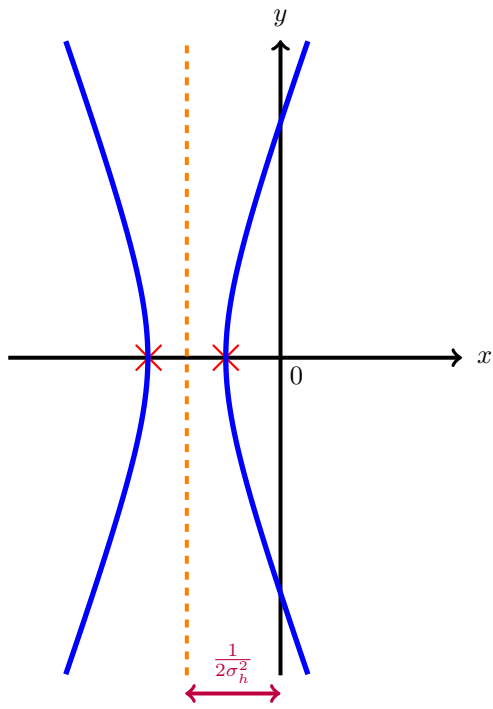


Figure 5.1: Illustration of the hyperbola described by equation (5.53).

Completing the square and dividing by $\frac{2}{\sigma_h^2}$ yields the equation of a hyperbola:

$$4\sigma_h^4(4\sigma_h^2+1)(2\sigma_h^2+1)\left(x+\frac{1}{2\sigma_h^2}\right)^2-\frac{\sigma_h^2}{2}(4\sigma_h^2+1)y^2=1\,. \tag{5.53}$$

This hyperbola is shown in Figure 5.1 for a specific choice of the parameter σ_h^2. Notice that this hyperbola is symmetric with respect to the x-axis and symmetric with respect to the vertical axis $x=-\frac{1}{2\sigma_h^2}$. Moreover, it holds that

$$x=-\frac{1}{2\sigma_h^2}\pm\frac{1}{2\sigma_h^2\sqrt{(1+4\sigma_h^2)(1+2\sigma_h^2)}},y=0$$

are the two solutions of equation (5.53) intersecting the x-axis. Hence, for values $s_2=x+\mathrm{i}y$ on the right-branch of this hyperbola it holds that

$$\mathrm{Re}[s_2]\geq-\frac{1}{2\sigma_h^2}+\frac{1}{2\sigma_h^2\sqrt{(1+4\sigma_h^2)(1+2\sigma_h^2)}}\,.$$

Since $\sigma_h^2\geq\frac{1}{4}$, a part of this branch is always located in the left half-plane. In particular, $s_2=0$ is located in the interior of the region bounded by the right-branch of the hyperbola.

We have thus shown that for all complex values z located on the hyperbola with equation (5.53), there exists a unique, positive real s_1 given by (5.52) such that $K_h(s_1, z) = K_h(s_1, \bar{z}) = 0$. However, in order to guarantee that $U_h(s_1, z)$ and $U_h(s_1, \bar{z})$ are finite, we restrict ourselves to the right-branch of the hyperbola. Because of (5.20), it holds that $-\frac{4}{1+4\sigma_h^2} < -\frac{1}{2\sigma_h^2}$. Consequently, by restricting to the right-branch of the hyperbola, $U_h(s_1, s_2)$ is certainly finite and we can safely use equation (5.48). Let us denote the right-branch of the hyperbola by Σ. More precisely, we define

$$\Sigma = \left\{ (x,y) \in \mathbb{R}^2 \mid 4\sigma_h^4(1+4\sigma_h^2)(1+2\sigma_h^2)\left(x + \frac{1}{2\sigma_h^2}\right)^2 - \frac{\sigma_h^2}{2}(4\sigma_h^2+1)y^2 = 1, x > -\frac{1}{2\sigma_h^2} \right\} . \quad (5.54)$$

The curve Σ divides the complex plane into two parts. The region onto the right of it, containing the origin, will in the remainder be referred to as the interior region of Σ.

For $\tilde{s}_2 = s$, $\hat{s}_2 = \bar{s}$, $s \in \Sigma$, equation (5.48) implies that

$$\left(\sigma_h^2 + \frac{1}{s}\right) S_h(s) = \left(\sigma_h^2 + \frac{1}{\bar{s}}\right) S_h(\bar{s}) ,$$

or, using that $S_h(\bar{s}) = \overline{S_h(s)}$,

$$\mathrm{Im}\left[\left(\sigma_h^2 + \frac{1}{s}\right) S_h(s)\right] = 0, \quad s \in \Sigma . \quad (5.55)$$

Observe that $\left(\sigma_h^2 + \frac{1}{s}\right) S_h(s)$ is analytic in the interior region of Σ, except at $s = 0$ where it has a simple pole. We thus have reduced our problem to that of determining a function that is analytic inside the interior region of Σ, except for a simple pole at $s = 0$, with prescribed boundary values of its imaginary part (5.55). The solution to this problem in case the boundary is the unit circle, is given in [49, Sect. I.3.3]. Hence, our boundary-value problem can be solved using a conformal mapping. Therefore let f_0 be a conformal mapping from the unit disk onto the interior of Σ. Since the real axis is an axis of symmetry of the interior of Σ, it is natural to choose (as we may) f_0 symmetric with respect to the real axis [7, Ch. VI.4], i.e. $f_0(\bar{s}) = \overline{f_0(s)}$, and hence that $f_0(1) = \infty$. We can then rewrite (5.55) as

$$\mathrm{Im}\left[\left(\sigma_h^2 + \frac{1}{f_0(s)}\right) S_h(f_0(s))\right] = 0, \quad |s| = 1 . \quad (5.56)$$

Finally, let us denote s_0 as the unique value in the open interval $]-1, 1[$ such that $f_0(s_0) = 0$. We are now facing the following problem: find a function

$\left(\sigma_h^2 + \frac{1}{f_0(s)}\right) S_h(f_0(s))$ analytic inside the unit disk, except for a simple pole at $s = s_0$, satisfying the boundary condition (5.56). The solution of this problem is given by : [49, Sect. I.3.3]

$$\left(\sigma_h^2 + \frac{1}{f_0(s)}\right) S_h(f_0(s)) = c_0 - \mathrm{i}c_1 \frac{s - s_0}{1 - ss_0} + \mathrm{i}\bar{c_1} \frac{1 - ss_0}{s - s_0} , \tag{5.57}$$

with c_0, c_1 two unknown constants that still have to be determined. Let $f = f_0^{-1}$ be the conformal mapping from the interior region of Σ to the unit disk. Then

$$\left(\sigma_h^2 + \frac{1}{s}\right) S_h(s) = c_0 - \mathrm{i}c_1 \frac{f(s) - f(0)}{1 - f(s)f(0)} + \mathrm{i}\bar{c_1} \frac{1 - f(s)f(0)}{f(s) - f(0)} , \tag{5.58}$$

or

$$S_h(s) = \left(c_0 - \mathrm{i}c_1 \frac{f(s) - f(0)}{1 - f(s)f(0)} + \mathrm{i}\bar{c_1} \frac{1 - f(s)f(0)}{f(s) - f(0)}\right) \frac{s}{1 + \sigma_h^2 s} . \tag{5.59}$$

Since $S_h(s)$ is the LST of a random variable, it is required that $S_h(0) = 1$. Taking the limit $s \to 0$ into the expression above yields

$$\begin{aligned} 1 &= \lim_{s\to 0} \left(c_0 - \mathrm{i}c_1 \frac{f(s) - f(0)}{1 - f(s)f(0)} + \mathrm{i}\bar{c_1} \frac{1 - f(s)f(0)}{f(s) - f(0)}\right) \frac{s}{1 + \sigma_h^2 s} \\ \Leftrightarrow 1 &= \lim_{s\to 0} \mathrm{i}\bar{c_1} \frac{1 - f(s)f(0)}{f(s) - f(0)} \frac{s}{1 + \sigma_h^2 s} \\ \Leftrightarrow 1 &= \mathrm{i}\bar{c_1} \frac{1 - f^2(0)}{f'(0)} . \end{aligned}$$

We thus see that $\mathrm{i}\bar{c_1} = \frac{f'(0)}{1 - f^2(0)}$, yielding

$$S_h(s) = \left(c_0 + \frac{f'(0)}{1 - f^2(0)} \left(\frac{f(s) - f(0)}{1 - f(s)f(0)} + \frac{1 - f(s)f(0)}{f(s) - f(0)}\right)\right) \frac{s}{1 + \sigma_h^2 s} .$$

Further, from the initial value theorem for LSTs, we must have that

$$\lim_{s\to\infty} sS_h(s)$$

is finite, which can only be if (using that $\lim_{s\to\infty} f(s) = 1$)

$$c_0 = -\frac{2f'(0)}{1 - f^2(0)} .$$

As a result,

$$S_h(s) = \frac{f'(0)}{1 - f^2(0)} \left(-2 + \frac{f(s) - f(0)}{1 - f(s)f(0)} + \frac{1 - f(s)f(0)}{f(s) - f(0)}\right) \frac{s}{1 + \sigma_h^2 s} . \tag{5.60}$$

The function f can be explictly obtained. The linear function $z_1 = z + \frac{1}{2\sigma_h^2}$ maps the region described by

$$4\sigma_h^4(1+4\sigma_h^2)(1+2\sigma_h^2)\left(x+\frac{1}{2\sigma_h^2}\right)^2 - \frac{\sigma_h^2}{2}(1+4\sigma_h^2)y^2 > 1\ ,$$

$z = x + \mathrm{i}y$, to the region

$$4\sigma_h^4(1+4\sigma_h^2)(1+2\sigma_h^2)x^2 - \frac{\sigma_h^2}{2}(1+4\sigma_h^2)y^2 > 1\ .$$

In [133] (page 186, Equation (13)), it is shown that the interior of the right branch of such a hyperbola is mapped to the upper half-plane by the function $z_2 = \mathrm{i}\sqrt{2}\cosh\left(\frac{\pi}{2\varphi}\mathrm{Arccosh}\frac{z_1}{k}\right)$, with

$$k = \frac{1}{2\sigma_h^2}\sqrt{\frac{1+4\sigma_h^2}{1+2\sigma_h^2}}\ , \tag{5.61}$$

$$\cos\varphi = \frac{1}{1+4\sigma_h^2}\ , \quad 0 \le \varphi < \frac{\pi}{2}\ . \tag{5.62}$$

Finally, the Möbius function $z_3 = -\frac{\sqrt{2}+\mathrm{i}z_2}{\sqrt{2}-\mathrm{i}z_2}$, maps the upper half-plane to the unit disk such that $\mathrm{i}\sqrt{2}$ is mapped to 0 and $-\mathrm{i}\sqrt{2}$ to infinity (see for example [7], page 175, Equation (17)) . The composition of these mappings gives us the mapping f:

$$f(s) = \tanh^2\left(\frac{\pi}{4\varphi}\mathrm{Arccosh}\left\{\frac{s}{k}+\cos\frac{\varphi}{2}\right\}\right)\ , \tag{5.63}$$

where we used that

$$\frac{1}{2\sigma_h^2 k} = \sqrt{\frac{1+2\sigma_h^2}{1+4\sigma_h^2}} = \cos\frac{\varphi}{2}\ . \tag{5.64}$$

Remark that due to the particular choice of the Möbius function z_3, the hyperbolic tangent tanh shows up. One can verify that the function $f(s)$ is indeed symmetric with respect to the x-axis. Further it is not difficult to obtain that

$$f(0) = \tanh^2\left(\frac{\mathrm{i}\pi}{8}\right) = -3+2\sqrt{2}\ . \tag{5.65}$$

Substituting $f(s)$ and $f(0)$ into (5.60) and using some trigonometric identities gives us the following final result for $S_h(s)$:

$$S_h(s) = \frac{\pi}{\varphi k\sin(\frac{\varphi}{2})}\frac{s}{\cosh\left(\frac{\pi}{\varphi}\mathrm{Arccosh}\left\{\frac{s}{k}+\cos\frac{\varphi}{2}\right\}\right)(\sigma_h^2 s+1)}\ . \tag{5.66}$$

The joint LST of the random variables $(1-2\lambda)u_1$ and $(1-2\lambda)u_2$, as $\lambda \uparrow \frac{1}{2}$ is then obtained by solving (5.43) for $U_h(s_1, s_2)$ and substituting (5.66).

We note that, despite the presence of the function Arccosh, $S_h(s)$ is indeed analytic in the right half-plane (as it should be for the LST of a continuous random variable). The function Arccosh is the analytic inverse of cosh : $\{z \in \mathbb{C} : 0 < \operatorname{Im} z < \pi\} \to \mathbb{C} \setminus \{a : a \in \mathbb{R}, |a| \geq 1\}$. Clearly, Arccosh is not continuous on the branch cut $[1, +\infty[$, since it holds that for $x \in [1, +\infty[$, $\lim_{z->x, \operatorname{Im} z<0} \operatorname{Arccosh} z = -\lim_{z->x, \operatorname{Im} z>0} \operatorname{Arccosh} z = -\operatorname{Arccosh} x$. But since $\cosh(-z) = \cosh(z)$, the limit of the composite transformation yields in both cases the same result. By the Riemann continuation theorem, the function S_h is indeed analytic in the right half-plane.

Finally, remark that $S_h(s)$ is the product of the LST of an exponential variable with mean σ_h^2 and the function

$$\frac{\pi}{\varphi k \sin(\frac{\varphi}{2})} \frac{s}{\cosh\left(\frac{\pi}{\varphi} \operatorname{Arccosh}\left\{\frac{s}{k} + \cos\frac{\varphi}{2}\right\}\right)},$$

which satisfies also the normalization property. It is not known to us if this function is the LST of a known distribution.

5.3.3 Calculation of moments

The determination of all (mixed) moments of the scaled system contents, as $\lambda \uparrow \frac{1}{2}$, can be deduced from the LST $U_h(s_1, s_2)$. In this section we are particularly interested in the correlation coefficient. First recall that we obtained the marginal LST of the scaled system contents already from the start, see (5.45), which was the LST of an exponential distribution. In particular, we have that

$$\lim_{\lambda \uparrow \frac{1}{2}} \mathsf{E}[(1-2\lambda)u_1] = \sigma_h^2 + \frac{1}{4} \tag{5.67}$$

$$\lim_{\lambda \uparrow \frac{1}{2}} \operatorname{var}[(1-2\lambda)u_1] = \left(\sigma_h^2 + \frac{1}{4}\right)^2 . \tag{5.68}$$

By symmetry, the mean and variance of $(1-2\lambda)u_2$ are equal to the mean and variance of $(1-2\lambda)u_1$. Therefore, we also obtain the mean of the total scaled system content

$$\lim_{\lambda \uparrow \frac{1}{2}} \mathsf{E}[(1-2\lambda)(u_1+u_2)] = 2\sigma_h^2 + \frac{1}{2} . \tag{5.69}$$

Note that we already obtained this result as (5.21) in Section 5.2. Also, note that this expression can be found by taking the first derivative of (5.66), substituting $s = 0$ and multiplying by -1. Making use of (5.66), we can compute higher moments of the scaled total system content using the moment-generating

property of LSTs. For the second moment, we get

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}[((1-2\lambda)(u_1+u_2))^2] = S_h''(0) = \frac{\left(4\sigma_h^4 - 10\sigma_h^2 - 3\right)\varphi^2 + 8\pi^2\sigma_h^4 + 4\pi^2\sigma_h^2}{6\varphi^2} . \quad (5.70)$$

Since for a random variable X it holds that $\mathrm{var}[X] = \mathsf{E}[X^2] - \mathsf{E}[X]^2$, it easily follows that

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathrm{var}[(1-2\lambda)(u_1+u_2)] = \frac{-\left(40\sigma_h^4 + 44\sigma_h^2 + 9\right)\varphi^2 + 8\pi^2\sigma_h^4 + 4\pi^2\sigma_h^2}{12\varphi^2} . \quad (5.71)$$

Since also for two random variables X, Y it holds that $\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}[X,Y]$, we get

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathrm{cov}[(1-2\lambda)u_1, (1-2\lambda)u_2] = \frac{-\left(128\sigma_h^4 + 112\sigma_h^2 + 21\right)\varphi^2 + 32\pi^2\sigma_h^4 + 16\pi^2\sigma_h^2}{48\varphi^2} . \quad (5.72)$$

The correlation between the system contents can then be found as

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathrm{corr}[u_1, u_2] = \lim_{\lambda\uparrow\frac{1}{2}} \mathrm{corr}[(1-2\lambda)u_1, (1-2\lambda)u_2] = \frac{32\pi^2\sigma_h^4 - 128\sigma_h^4\varphi^2 + 16\pi^2\sigma_h^2 - 112\sigma_h^2\varphi^2 - 21\varphi^2}{3\varphi^2\left(1+4\sigma_h^2\right)^2} . \quad (5.73)$$

5.3.4 Examples and discussions

The results obtained in the previous subsection depend only on the system parameter σ_h^2. Note that φ is given by

$$\varphi = \arccos\left(\frac{1}{1+4\sigma_h^2}\right) .$$

Hence, if σ_h^2 ranges from $\frac{1}{4}$ to $+\infty$, it follows that φ ranges from

$$\frac{\pi}{3} \le \varphi \le \frac{\pi}{2} .$$

We emphasize that the results from the previous subsection are valid for any arrival distribution with pgf $A(z)$, provided that $A''(1)$ has a finite limit for $\lambda \uparrow \frac{1}{2}$. Before discussing the results in general, we first treat the two extreme cases for σ_h^2, i.e. $\sigma_h^2 = \frac{1}{4}$ and $\sigma_h^2 \to +\infty$.

5.3.4.1 Bernoulli arrivals

Let us assume that the number of type-j arrivals within a slot is Bernoulli distributed, i.e.

$$A(z) = 1 - \lambda + \lambda z \ .$$

Since $A''(1) = 0$, the lower bound (5.20) for σ_h^2 is an equality, i.e.

$$\sigma_h^2 = \frac{1}{4} \ .$$

The parameter φ defined in (5.62) is then given by

$$\varphi = \frac{\pi}{3} \ .$$

Using the fact that $\cosh(3z) = 4\cosh^3(z) + 3\cosh(z)$ for every z, we obtain that the LST $S_h(s)$ simplifies to

$$S_h(s) = \frac{32}{(s+2)(s+4)^2} \ .$$

We thus observe that, in case of Bernoulli arrivals, $S_h(s)$ is the product of three LSTs of exponential random variables. This simple expression is completely in accordance with the results obtained in Section 2.3. In Section 2.3, we studied the non-work-conserving model under the assumption of (not necessarily symmetric) Bernoulli arrivals. For this particular case of arrivals, we obtained the joint probability distribution of the number of type-1 and type-2 customers, in steady state. The corresponding joint probability generating function turned out to be a rational function.

The correlation coefficient between the type-1 and type-2 customers, using (5.73), is equal to

$$\lim_{\lambda \uparrow \frac{1}{2}} \mathrm{corr}[u_1, u_2] = -\frac{1}{4} \ .$$

5.3.4.2 Arrivals with infinite asymptotic variance

We have made the restriction that $A''(1)$ has a finite limit for $\lambda \uparrow \frac{1}{2}$. As a consequence, this implies that σ_h^2 remains finite. However, in the correlation coefficient (5.73) we still can take the limit as $\sigma_h^2 \to +\infty$. For this limit, we have that $\varphi \to \frac{\pi}{2}$. Using (5.73), we obtain that

$$\lim_{\lambda \uparrow \frac{1}{2}} \mathrm{corr}[u_1, u_2] \to 0, \quad \text{as } \sigma_h^2 \to +\infty \ .$$

This result suggests that in case the mean and the variance of the number of arrivals is very high in both queues, the two queues become uncorrelated.

5.3.4.3 Other arrival processes

For other well-known arrival processes, such as the geometric distribution, binomial distribution and the Poisson distribution, the LST $S_h(s)$ does not simplify to a rational function. Nevertheless, for every value of $\sigma_h^2 \in \left[\frac{1}{4}, +\infty\right]$, we can easily compute the correlation coefficient using (5.73). In Figure 5.2, we show the correlation coefficient for $\sigma_h^2 \in \left[\frac{1}{4}, 10\right]$, which quantifies the correlation between the number of type-1 and type-2 customers in the system at the beginning of a slot for $\lambda \uparrow \frac{1}{2}$.

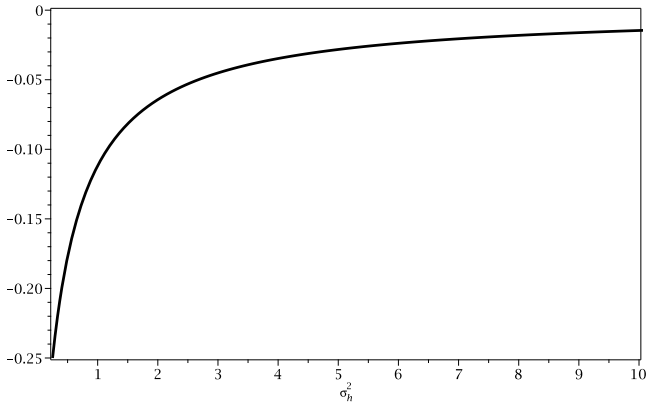


Figure 5.2: Correlation coefficient between the system contents for $\lambda \uparrow \frac{1}{2}$, versus the asymptotic arrival variance σ_h^2.

Looking at Figure 5.2, we observe that the correlation coefficient is always negative. It is worth noting that in Section 2.3, it has been shown that for the special case of two independent Bernoulli arrival processes, the correlation coefficient is negative for *all* allowed values of λ. The reason for this negative correlation coefficient is the service policy. If a queue is getting longer, this is because either there were a lot of arrivals in the previous slots or because the queue was not getting served the previous slots. So in the latter case, if queue-1 is large, it is likely that queue-2 is small. In the heavy-traffic limit $\lambda \uparrow \frac{1}{2}$, the lengths of both queues go to infinity. However, according to Figure 5.2, the correlation effect of the service policy is still present for σ_h^2 not too high. Finally, we observe that the correlation coefficient is a strictly increasing function of σ_h^2 and goes from $-\frac{1}{4}$ to 0 as σ_h^2 increases from $\frac{1}{4}$ to $+\infty$.

5.4 The work-conserving policy in heavy-traffic

We now analyze the joint distribution of v_1 and v_2 in case of heavy traffic. More concretely, the purpose is to obtain the joint LST of the scaled random variables $(1-2\lambda)v_1$ and $(1-2\lambda)v_2$ as $\lambda \uparrow \frac{1}{2}$. We assume that this limit distribution

exists and we define the corresponding LST of this distribution as

$$V_h(s_1, s_2) \triangleq \lim_{\lambda \uparrow \frac{1}{2}} \mathsf{E}[e^{-s_1(1-2\lambda)v_1 - s_2(1-2\lambda)v_2}] \,. \tag{5.74}$$

The subscript h is used again to indicate that the LST corresponds to the heavy-traffic limit. Similarly we define

$$\begin{aligned} Q_h(s) &\triangleq \lim_{\lambda \uparrow \frac{1}{2}} \mathsf{E}[e^{-s(1-2\lambda)v_1}] \\ &= V_h(s, 0) \end{aligned} \tag{5.75}$$

as the limiting LST of the number of type-1 customers in the system. Obviously because of the symmetry we also have

$$\begin{aligned} Q_h(s) &= \lim_{\lambda \uparrow \frac{1}{2}} \mathsf{E}[e^{-s(1-2\lambda)v_2}] \\ &= V_h(0, s) \end{aligned}$$

as the limiting LST of the number of type-2 customers in the system. To obtain a functional equation for $V_h(s_1, s_2)$, we can substitute $z_1 = e^{-s_1(1-2\lambda)}$ and $z_2 = e^{-s_2(1-2\lambda)}$ into (5.12) and take the limit $\lambda \uparrow 1/2$. Let us first rewrite the boundary function $V(e^{-(1-2\lambda)s}, 0)$ as a function of $Q_h(s)$. To that end, we solve equation (5.16) for $V(z, 0) - V(0, 1)$. This gives us

$$V(z, 0) - V(0, 1) = 1 - 2\lambda - \frac{2z - (z+1)A(z)}{A(z)(z-1)} V(z, 1) \,.$$

Next, we divide both sides by $1 - 2\lambda$ and use the change of variable $z = e^{-s(1-2\lambda)}$. Finally, taking the limit $\lambda \uparrow 1/2$ we obtain that

$$\lim_{\lambda \uparrow 1/2} \frac{1}{1-2\lambda} \left(V(e^{-s(1/2-\lambda)}, 0) - V(0, 1) \right) = 1 - \left(\left(\sigma_h^2 + \frac{1}{4} \right) s + 1 \right) Q_h(s) \,. \tag{5.76}$$

The detailed computations to obtain the limit above are omitted. The computations are very similar to the ones in Section 5.3 to obtain equation (5.42), i.e. by carefully taking into account the dependency of $A(z)$ on the arrival rate λ and applying l'Hôpital's rule several times.

A functional equation for $V_h(s_1, s_2)$ can now be obtained, similarly as in Section 5.3. Rewrite (5.12) as

$$\begin{aligned} \frac{K(z_1, z_2)}{(z_2 - z_1)(1-2\lambda)} V(z_1, z_2) = A(z_1)A(z_2) \frac{1}{2} \Bigg(& \frac{1}{1-2\lambda} \left(V(z_1, 0) - V(1, 0) \right) \\ & - \frac{1}{1-2\lambda} \left(V(z_2, 0) - V(1, 0) \right) + \frac{2z_1 z_2 - z_1 - z_2}{z_2 - z_1} \Bigg) \,. \end{aligned} \tag{5.77}$$

Next, we substitute $z_1 = e^{-s_1(1/2-\lambda)}$ and $z_2 = e^{-s_2(1/2-\lambda)}$ into the above. Finally, taking the limit $\lambda \uparrow \frac{1}{2}$, we obtain the following functional equation for $V_h(s_1, s_2)$

$$\begin{aligned}\frac{K_h(s_1,s_2)}{4(s_2-s_1)}V_h(s_1,s_2) = &\left(\left(\sigma_h^2+\frac{1}{4}\right)s_2+1\right)Q_h(s_2)\\ &-\left(\left(\sigma_h^2+\frac{1}{4}\right)s_1+1\right)Q_h(s_1)+\frac{s_1+s_2}{s_2-s_1}\,.\end{aligned} \tag{5.78}$$

The LST of the total scaled system content can easily be obtained by substituting $s_1 = s_2 = s$ in equation (5.78), yielding

$$V_h(s,s) = \frac{1}{\sigma_h^2 s+1}\,. \tag{5.79}$$

The above LST is the LST of an exponentially distributed random variable with mean σ_h^2.

5.4.1 Solution of the functional equation

In this section, we will determine the boundary function $Q_h(s)$, and hence the solution $V_h(s_1, s_2)$ of the functional equation (5.78). We solve equation (5.78) in the same manner as we did in Section 5.3, i.e. using the boundedness of the function $V_h(s_1, s_2)$ and the zeros of $K_h(s_1, s_2)$ in order to obtain a boundary-value problem for the remaining unknown function $Q_h(s)$. Regarding the boundedness, we observe from (5.79) that $V_h(s_1, s_2)$ is analytic for $\mathrm{Re}[s_1] = \mathrm{Re}[s_2] > -\frac{1}{\sigma_h^2}$. Hence, since $V_h(s_1, s_2)$ is a joint LST it follows that $V_h(s_1, s_2)$ is also analytic for

$$\mathrm{Re}[s_1] > \mathrm{Re}[s_2] > -\frac{1}{\sigma_h^2} \quad \text{and} \quad \mathrm{Re}[s_2] > \mathrm{Re}[s_1] > -\frac{1}{\sigma_h^2}\,. \tag{5.80}$$

In particular, the marginal LST $Q_h(s) = V_h(s, 0)$ is analytic for at least $\mathrm{Re}[s] > -\frac{1}{\sigma_h^2}$.

We now proceed as in Section 5.3. For any given value s_1 there exist exactly two values, say $\tilde{s}_2$ and $\hat{s}_2$, such that $K_h(s_1, \tilde{s}_2) = 0$, $K_h(s_1, \hat{s}_2) = 0$. If moreover $V_h(s_1, \tilde{s}_2)$ and $V_h(s_1, \hat{s}_2)$ are finite, we obtain from (5.78) that

$$\left(\left(\sigma_h^2+\frac{1}{4}\right)\tilde{s}_2+1\right)Q_h(\tilde{s}_2) - \left(\left(\sigma_h^2+\frac{1}{4}\right)s_1+1\right)Q_h(s_1)+\frac{s_1+\tilde{s}_2}{\tilde{s}_2-s_1} = 0\,,$$

and

$$\left(\left(\sigma_h^2+\frac{1}{4}\right)\hat{s}_2+1\right)Q_h(\hat{s}_2) - \left(\left(\sigma_h^2+\frac{1}{4}\right)s_1+1\right)Q_h(s_1)+\frac{s_1+\hat{s}_2}{\hat{s}_2-s_1} = 0\,,$$

so that by eliminating $Q_h(s_1)$ we find

$$\left(\left(\sigma_h^2+\frac{1}{4}\right)\tilde{s}_2+1\right)Q_h(\tilde{s}_2)-\left(\left(\sigma_h^2+\frac{1}{4}\right)\hat{s}_2+1\right)Q_h(\hat{s}_2)=\frac{s_1+\hat{s}_2}{\hat{s}_2-s_1}-\frac{s_1+\tilde{s}_2}{\tilde{s}_2-s_1}\,. \tag{5.81}$$

We emphasize that the kernel $K_h(s_1,s_2)$ in Equation (5.78) is the same as in Section 5.3. The analysis of the kernel in Section 5.3 can thus be applied in this section as well. For every $s_2\in\Sigma$, defined by (5.54), there exists a unique, positive real s_1 such that $K_h(s_1,s_2)=K_h(s_1,\bar{s_2})=0$. Since for every $s_2\in\Sigma$ it holds that $\mathrm{Re}[s_2]>-\frac{1}{2\sigma_h^2}>-\frac{1}{\sigma_h^2}$, $V_h(s_1,s_2)$ is bounded whenever $s_2\in\Sigma$ and when s_1 is given by (5.52). Then, equation (5.81) implies that for $\tilde{s}_2=s$, $\hat{s}_2=\bar{s}$, $s\in\Sigma$, (with $x=\mathrm{Re}[s]$, $y=\mathrm{Im}[s]$):

$$\left(\left(\sigma_h^2+\frac{1}{4}\right)s+1\right)Q_h(s)-\left(\left(\sigma_h^2+\frac{1}{4}\right)\bar{s}+1\right)Q_h(\bar{s})=\frac{4s_1\mathrm{i}y}{x^2+y^2-2s_1x+s_1^2}\,,$$

where s_1 is given by (5.52). Substituting (5.52) and (5.50) into above gives us

$$\left(\left(\sigma_h^2+\frac{1}{4}\right)s+1\right)Q_h(s)-\left(\left(\sigma_h^2+\frac{1}{4}\right)\bar{s}+1\right)Q_h(\bar{s})=\frac{\mathrm{i}y}{2\frac{1+2\sigma_h^2}{1+4\sigma_h^2}+2\sigma_h^2x}\,.$$

Since $\overline{Q_h(s)}=Q_h(\bar{s})$, we have obtained that

$$\mathrm{Im}\left[\left(\left(\sigma_h^2+\frac{1}{4}\right)s+1\right)Q_h(s)\right]=\frac{\mathrm{Im}[s]}{4\frac{1+2\sigma_h^2}{1+4\sigma_h^2}+4\sigma_h^2\,\mathrm{Re}[s]},\quad s\in\Sigma\,. \tag{5.82}$$

Let f_0 again be a conformal map from the unit disk to interior region of Σ, such that f_0 is symmetric with respect to the real axis. We can thus write that

$$\mathrm{Im}\left[\left(\left(\sigma_h^2+\frac{1}{4}\right)f_0(s)+1\right)Q_h(f_0(s))\right]=\frac{\mathrm{Im}[f_0(s)]}{4\frac{1+2\sigma_h^2}{1+4\sigma_h^2}+4\sigma_h^2\,\mathrm{Re}[f_0(s)]},\quad |s|=1\,. \tag{5.83}$$

The solution of this boundary-value problem is given by Schwarz integral formula [96, Th. 7.38], yielding

$$\left(\left(\sigma_h^2+\frac{1}{4}\right)f_0(s)+1\right)Q_h(f_0(s))=\frac{1}{2\pi}\int_{C_0}\frac{\zeta+s}{\zeta-s}v(f_0(\zeta))\frac{d\zeta}{\zeta}+D,\quad |s|<1\,, \tag{5.84}$$

where C_0 is the positively oriented unit circle, D a normalization constant constant and

$$v(s)=\frac{\mathrm{Im}[s]}{4\frac{1+2\sigma_h^2}{1+4\sigma_h^2}+4\sigma_h^2\,\mathrm{Re}[s]}\,. \tag{5.85}$$

Finally, let $f = f_0^{-1}$ be the inverse mapping of f_0. Then $Q_h(s)$ is given by

$$Q_h(s) = \left(\left(\sigma_h^2 + \frac{1}{4}\right)s + 1\right)^{-1} \left(\frac{1}{2\pi}\int_{C_0} \frac{\zeta + f(s)}{\zeta - f(s)} v(f_0(\zeta)) \frac{d\zeta}{\zeta} + D\right) , \quad (5.86)$$

for all s in the interior of Σ. The function f is given by (5.63). We emphasize that $s = 0$ is located in this region, hence we can obtain integral formulas from (5.86) for all derivatives of $Q_h(s)$ evaluated at $s = 0$. The joint LST $V_h(s_1, s_2)$ is fully determined by substituting (5.86) into (5.78).

In the remainder of this section, we will rewrite the contour integral in (5.86) as a real integral that is suitable for numerical integration.

5.4.2 Rewriting the contour integral in (5.86) as a real integral

Using transformation of contour integrals, we can write

$$\frac{1}{2\pi}\int_{C_0} \frac{\zeta + s}{\zeta - s} v(f_0(\zeta)) \frac{d\zeta}{\zeta} = \frac{1}{2\pi}\int_{\Sigma} \frac{f(z) + f(s)}{f(z) - f(s)} v(z) \frac{f'(z)dz}{f(z)} . \quad (5.87)$$

The most natural parametrization $z(t)$, $t \in \mathbb{R}$, of the contour Σ is given by

$$\begin{aligned} z(t) &= k\cos\varphi\cosh(t) - k\cos\frac{\varphi}{2} + \mathrm{i}k\sin\varphi\sinh(t) \\ &= k\cosh(t + \mathrm{i}\varphi) - k\cos\frac{\varphi}{2} . \end{aligned} \quad (5.88)$$

Notice that using this parametrization, the contour Σ moves clockwise around $s = 0$. Further, straightforward calculations yield

$$f(z(t)) = \tanh^2\left(\frac{\pi}{4\varphi}t + \frac{\mathrm{i}\pi}{4}\right) , \quad (5.89)$$

$$f'(z(t))z'(t) = \frac{\pi}{2\varphi}\frac{\tanh\left(\frac{\pi}{4\varphi}t + \frac{\mathrm{i}\pi}{4}\right)}{\cosh^2\left(\frac{\pi}{4\varphi}t + \frac{\mathrm{i}\pi}{4}\right)} , \quad (5.90)$$

$$\begin{aligned} v(z(t)) &= \frac{k\sin(\varphi)\sinh(t)}{4\cos^2\left(\frac{\varphi}{2}\right) + 4\sigma_h^2(k\cos(\varphi)\cosh(t) - k\cos\left(\frac{\varphi}{2}\right))} \\ &= \frac{k\tan(\varphi)\sinh(t)}{2 + 4\sigma_h^2 k\cosh(t)} . \end{aligned} \quad (5.91)$$

With the parametrization (5.88), the contour integral in (5.87) becomes

$$
\begin{aligned}
&\frac{1}{2\pi}\int_{\Sigma}\frac{f(z)+f(s)}{f(z)-f(s)}v(z)\frac{f'(z)dz}{f(z)}\\
&=-\frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{f(z(t))+f(s)}{f(z(t))-f(s)}v(z(t))\frac{f'(z(t))z'(t)dt}{f(z(t))}\\
&=-\frac{k\tan(\varphi)}{4\varphi}\int_{-\infty}^{+\infty}\frac{f(z(t))+f(s)}{f(z(t))-f(s)}\frac{\sinh(t)}{\sinh\left(\frac{\pi}{2\varphi}t+\frac{\mathrm{i}\pi}{2}\right)(1+2\cosh(t)k\sigma_h^2)}dt\\
&=\frac{\mathrm{i}\,k\tan(\varphi)}{4\varphi}\int_{-\infty}^{+\infty}\frac{f(z(t))+f(s)}{f(z(t))-f(s)}\frac{\sinh(t)}{\cosh\left(\frac{\pi}{2\varphi}t\right)(1+2\cosh(t)k\sigma_h^2)}dt\,. \qquad (5.92)
\end{aligned}
$$

The minus sign in the first equality is because the parametrization $z(t)$ for Σ is negatively oriented. The second equality follows by substituting (5.89), (5.90), (5.91). Observe that

$$
\frac{\sinh(t)}{\cosh\left(\frac{\pi}{2\varphi}t\right)(1+2\cosh(t)k\sigma_h^2)} \qquad (5.93)
$$

is an odd function in $t\in\mathbb{R}$. We will now write

$$
\frac{f(z(t))+f(s)}{f(z(t))-f(s)}
$$

as the sum of an even and an odd function. To accomplish this, we multiply the expression above by

$$
\frac{\overline{f(z(t))}-f(s)}{\overline{f(z(t))}-f(s)}
$$

Using that $|f(z(t))|=1$,

$$
\begin{aligned}
\frac{f(z(t))+f(s)}{f(z(t))-f(s)}=&\frac{1-f^2(s)-\mathrm{i}2\,\mathrm{Im}[f(z(t))]f(s)}{1+f^2(s)-2f(s)\,\mathrm{Re}[f(z(t))]}\\
=&\frac{1-f^2(s)}{1+f^2(s)-2f(s)(2\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1)}\\
&-\mathrm{i}\frac{4\tanh\left(\frac{\pi}{2\varphi}t\right)\mathrm{sech}\left(\frac{\pi}{2\varphi}t\right)f(s)}{1+f^2(s)-2f(s)(2\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1)}
\end{aligned}
$$

and we notice that the first term of the right-hand-side is an even function in $t\in\mathbb{R}$, while the second term is an odd function in $t\in\mathbb{R}$. Therefore the first term does not contribute in the integral of (5.92) and we obtain that the expression (5.92) simplifies to

$$
\frac{2k\tan(\varphi)}{\varphi}\int_0^{+\infty}\chi_1(s,t)\chi_2(t)dt\,, \qquad (5.94)
$$

where

$$\chi_1(s,t) = \frac{f(s)}{(1+f(s))^2 - 4f(s)\tanh^2\left(\frac{\pi}{2\varphi}t\right)}\,, \tag{5.95}$$

$$\chi_2(t) = \frac{\tanh\left(\frac{\pi}{2\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)\left(1+2\cosh(t)k\sigma_h^2\right)}\,. \tag{5.96}$$

We have thus obtained that $Q_h(s)$ is given by

$$Q_h(s) = \left(\left(\sigma_h^2 + \frac{1}{4}\right)s + 1\right)^{-1}\left(\frac{2k\tan(\varphi)}{\varphi}\int_0^{+\infty}\chi_1(s,t)\chi_2(t) + D\right)dt\,. \tag{5.97}$$

The constant D can be obtained from the normalization condition $Q_h(0) = 1$, yielding

$$D = 1 + \frac{k\tan(\varphi)}{4\varphi}\int_0^{+\infty}\frac{\tanh\left(\frac{\pi}{\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)\left(1+2\cosh(t)k\sigma_h^2\right)}dt\,, \tag{5.98}$$

where we have used (5.65).

5.4.3 Calculation of moments

In this section we will compute the mean and the variance of the number of type-j customers ($j = 1, 2$) and the mean and variance of the total number of customers, as $\lambda \uparrow \frac{1}{2}$. From these performance measures, the covariance and correlation between the number of type-1 and type-2 customers can be deduced.

We commence with the total number of customers. Notice that we have obtained the LST of the total number of customers (5.79) without much effort. More precisely, (5.79) is the LST of an exponential distribution with mean σ_h^2. Hence $(1-2\lambda)(v_1+v_2)$ is exponentially distributed with mean σ_h^2 for $\lambda \uparrow \frac{1}{2}$. We can thus write that

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}\left[(1-2\lambda)(v_1+v_2)\right] = \sigma_h^2\,; \tag{5.99}$$

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathrm{var}\left[(1-2\lambda)(v_1+v_2)\right] = \sigma_h^4\,. \tag{5.100}$$

Since we are considering a symmetric system, v_1 and v_2 have the same distribution. Therefore, the scaled random variables $(1-2\lambda)v_1$ and $(1-2\lambda)v_2$ must have the same mean for $\lambda \uparrow \frac{1}{2}$. As a consequence, we thus also obtain the mean number of type-1 and type-2 customers:

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}\left[(1-2\lambda)v_1\right] = \lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}\left[(1-2\lambda)v_2\right] = \frac{\sigma_h^2}{2}\,. \tag{5.101}$$

Next, we will compute $\mathrm{var}[(1-2\lambda)v_1]$ as $\lambda \uparrow \frac{1}{2}$. The moments of $(1-2\lambda)v_1$, as $\lambda \uparrow \frac{1}{2}$, are determined by the derivatives of $Q_h(s)$ at $s=0$. Write

$$Q_h(s) = \left(\left(\sigma_h^2 + \frac{1}{4}\right)s + 1\right)^{-1} I(s)\ , \tag{5.102}$$

where

$$I(s) = \frac{2k\tan(\varphi)}{\varphi}\int_0^{+\infty} \chi_1(s,t)\chi_2(t) + D\ . \tag{5.103}$$

Differentiating (5.102) and using $I(0)=1$ gives

$$Q_h'(0) = -\sigma_h^2 - \frac{1}{4} + I'(0) \tag{5.104}$$

$$Q_h''(0) = 2\left(\sigma_h^2 + \frac{1}{4}\right)^2 - 2\left(\sigma_h^2 + \frac{1}{4}\right)I'(0) + I''(0)\ . \tag{5.105}$$

Using (5.101) we find that $Q_h'(0) = -\frac{\sigma_h^2}{2}$, and from (5.104) it must be that

$$I'(0) = \frac{\sigma_h^2}{2} + \frac{1}{4}\ . \tag{5.106}$$

Differentiating (5.103) twice with respect to s and substituting $s=0$ yields

$$I''(0) = \int_0^{+\infty} \frac{r(t)}{(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1)^2} \cdot \frac{\tanh\left(\frac{\pi}{\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)(1+2\cosh(t)k\sigma_h^2)}dt\ , \tag{5.107}$$

with

$$\begin{aligned} r(t) = {} & \frac{\pi}{\varphi^2}(1+2\sigma_h^2)^{\frac{3}{2}}\sqrt{1+4\sigma_h^2}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1\right) \\ & + \frac{\pi^2}{\varphi^3}2\sqrt{2}\sigma_h(1+2\sigma_h^2)\sqrt{1+4\sigma_h^2}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1\right)\ . \end{aligned} \tag{5.108}$$

For ease of notation, let us denote the integral (5.107) as J, i.e.

$$J \triangleq \int_0^{+\infty} \frac{r(t)}{(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1)^2} \cdot \frac{\tanh\left(\frac{\pi}{\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)(1+2\cosh(t)k\sigma_h^2)}dt\ . \tag{5.109}$$

From (5.105) it follows that

$$\lim_{\lambda\uparrow\frac{1}{2}} \mathsf{E}\left[\left[((1-2\lambda)v_1)^2\right]\right] = Q_h''(0) = \sigma_h^4 + \frac{1}{4}\sigma_h^2 + J\ . \tag{5.110}$$

Further, we can compute the variance

$$\lim_{\lambda\uparrow\frac{1}{2}} \operatorname{var}\left[(1-2\lambda)v_1\right] = \frac{3}{4}\sigma_h^4 + \frac{1}{4}\sigma_h^2 + J\,, \tag{5.111}$$

and the covariance

$$\lim_{\lambda\uparrow\frac{1}{2}} \operatorname{cov}[(1-2\lambda)v_1, (1-2\lambda)v_2] = -\frac{1}{4}\sigma_h^4 - \frac{1}{4}\sigma_h^2 - J\,. \tag{5.112}$$

Finally, the correlation coefficient is given by

$$\begin{aligned}\lim_{\lambda\uparrow\frac{1}{2}} \operatorname{corr}[v_1, v_2] &= \lim_{\lambda\uparrow\frac{1}{2}} \operatorname{corr}[(1-2\lambda)v_1, (1-2\lambda)v_2] \\ &= \frac{\sigma_h^4}{2}\left(\frac{3}{4}\sigma_h^4 + \frac{1}{4}\sigma_h^2 + J\right)^{-1} - 1\,. \end{aligned} \tag{5.113}$$

5.4.4 Examples and discussion

The only parameter that is present in the results that we have obtained in the previous subsection is the asymptotic variance of the number of arrivals per slot, i.e. σ_h^2. We recall that if σ_h^2 ranges from $\frac{1}{4}$ to $+\infty$, it follows that φ ranges from $\frac{\pi}{3} \leq \varphi \leq \frac{\pi}{2}$.

5.4.4.1 Arrivals with infinite asymptotic variance

Let us rewrite (5.113) as

$$\lim_{\lambda\uparrow\frac{1}{2}} \operatorname{corr}[v_1, v_2] = \frac{1}{2}\left(\frac{3}{4} + \frac{1}{4\sigma_h^2} + \frac{J}{\sigma_h^4}\right)^{-1} - 1$$

We are interested in the case that $\sigma_h^2 \to +\infty$. To that end, we have to compute the limit

$$\lim_{\sigma_h^2\to+\infty} \frac{J}{\sigma_h^4}\,.$$

For ease of notation, we make the change of variables $\mu = \sigma_h^2$. Thus, we write

$$\begin{aligned}&\lim_{\mu\to+\infty} \frac{J}{\mu^2} \\ &= \lim_{\mu\to+\infty} \int_0^{+\infty} \frac{1}{\mu^2} \frac{r_\mu(t)}{(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1)^2} \frac{\tanh\left(\frac{\pi}{\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)(1+2\cosh(t)k\mu)} dt\,,\end{aligned} \tag{5.114}$$

with

$$\begin{aligned} r_\mu(t) =& \frac{\pi}{\varphi^2}(1+2\mu)^{\frac{3}{2}}\sqrt{1+4\mu}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1\right) \\ &+\frac{\pi^2}{\varphi^3}2\sqrt{2\mu}(1+2\mu)\sqrt{1+4\mu}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1\right) \\ =& \frac{\pi}{\varphi^3}(1+2\mu)\sqrt{1+4\mu}\left(\varphi\sqrt{1+2\mu}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1\right)\right. \\ &\left.+\pi 2\sqrt{2\mu}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1\right)\right) . \end{aligned}$$

We now consider the limit of $\frac{J}{\mu^2}$ as $\mu \to +\infty$. We have that

$$\begin{aligned} \varphi &\to \frac{\pi}{2} \\ k &\to 0 \\ k\mu &\to \frac{\sqrt{2}}{2} , \end{aligned}$$

as $\mu \to +\infty$, cf. (5.61), (5.62). Hence,

$$\begin{aligned} &\lim_{\mu\to+\infty} \frac{1}{(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1)^2}\,\frac{\tanh\left(\frac{\pi}{\varphi}t\right)\sinh(t)}{\cosh^2\left(\frac{\pi}{2\varphi}t\right)(1+2\cosh(t)k\mu)} \\ &= \frac{\tanh(2t)\sinh(t)}{(\tanh^2(t)+1)^2\cosh^2(t)\left(1+\sqrt{2}\cosh(t)\right)} \end{aligned} \tag{5.115}$$

Next, we consider the limit of $\frac{r_\mu(t)}{\mu^2}$ as $\mu \to +\infty$. We write

$$\begin{aligned} \frac{r_\mu(t)}{\mu^2} =& \frac{\pi}{\varphi^3}\left(\frac{1}{\mu}+2\right)\sqrt{\frac{1}{\mu}+4}\left(\varphi\sqrt{\frac{1}{\mu}+2}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)+1\right)\right. \\ &\left.+\pi 2\sqrt{2}\left(\tanh^2\left(\frac{\pi}{2\varphi}t\right)-1\right)\right) . \end{aligned}$$

Whence,

$$\lim_{\mu\to+\infty}\frac{r_\mu(t)}{\mu^2} = \frac{16\sqrt{2}}{\pi}\left(5\tanh^2(t)-3\right) . \tag{5.116}$$

Due to (5.115), (5.116), and the fact that the absolute value of the integrand in (5.114) is bounded by the integrable function

$$C\frac{\sinh(t)}{\cosh^2(t)}, \quad \text{where } C \text{ is a positive constant} ,$$

we can apply Lebesgue's dominated convergence theorem to pass to the limit on the right-hand side in (5.114) and to obtain

$$\lim_{\mu\to+\infty}\frac{J}{\mu^2} = \frac{16\sqrt{2}}{\pi}\int_0^{+\infty}\frac{\left(5\tanh^2(t)-3\right)\tanh(2t)\sinh(t)}{(\tanh^2(t)+1)^2\cosh^2(t)\left(1+\sqrt{2}\cosh(t)\right)}dt$$
$$= 3-\frac{32}{3\pi}\,. \tag{5.117}$$

Thus, we have

$$\lim_{\lambda\uparrow\frac{1}{2}}\mathrm{corr}[v_1,v_2] \to \frac{128-39\pi}{45\pi-128} \approx 0.409664\,, \quad \text{as } \sigma_h^2\to+\infty\,. \tag{5.118}$$

5.4.4.2 Other arrival processes

To the best of our knowledge, the integral (5.109) cannot be calculated analytically for $\frac{\pi}{3}\leq\varphi<\frac{\pi}{2}$. Even for specific values of σ_h^2, such as $\sigma_h^2=\frac{1}{4}$, it seems to be unfeasible to calculate (5.109) analytically. The formulas for the moments in the work-conserving case are therefore visually more complicated, compared to those in the non-work-conserving case. However, since an explicit expression for the integrand in (5.109) is available, the integral can be determined numerically without much difficulties. The examples presented in this chapter were all obtained by applying the substitution $x=\tanh(\frac{\pi}{2\varphi}t)$ into (5.109) such that the integration interval is $[0,1]$ and no truncation procedure is needed. This integral is then approximated by repeated application of the trapezoidal rule [134, Ch. 9], in which we partition the interval $[0,1]$ into 250 equal subintervals.

In Figure 5.3, we show the correlation coefficient between the system contents for both the work-conserving-service policy and the non-work-conserving service policy. We emphasize that we already discussed the correlation coefficient for the non-work-conserving service policy in Section 5.3.4. In the case of the work-conserving policy, we see from Figure 5.3, that the correlation coefficient between the system contents is always positive, at least for $\lambda\uparrow\frac{1}{2}$. This result reveals that the correlation structure between the system contents in case of the work-conserving policy differs significantly to that of the non-work-conserving policy, even for $\lambda\uparrow\frac{1}{2}$.

The difference is quite remarkable since for $\lambda\uparrow\frac{1}{2}$, more and more customers are being queued. We would have thought that, with the non-work conserving policy, the server is always allocated to a non-empty queue if $\lambda\uparrow\frac{1}{2}$ (just like with the work-conserving policy). However, a simple mean value analysis (5.23) already reveals that the system contents must be different for $\lambda\uparrow\frac{1}{2}$, since in this case,

$$\mathsf{E}[u_1+u_2]-\mathsf{E}[v_1+v_2] \sim \frac{\frac{1}{2}+\sigma_h^2}{1-2\lambda}\,,$$

cf. (5.23), which is nonzero. To strengthen this observation, we compare the correlation coefficient and use the fact that the correlation coefficient between

the scaled system contents equals the correlation coefficient between the system contents, cf. (5.17). Regarding the great difference in Figure 5.3, we can conclude that both stochastic processes, i.e. the non-work-conserving policy and the work-conserving policy, are considerably different for $\lambda \uparrow \frac{1}{2}$.

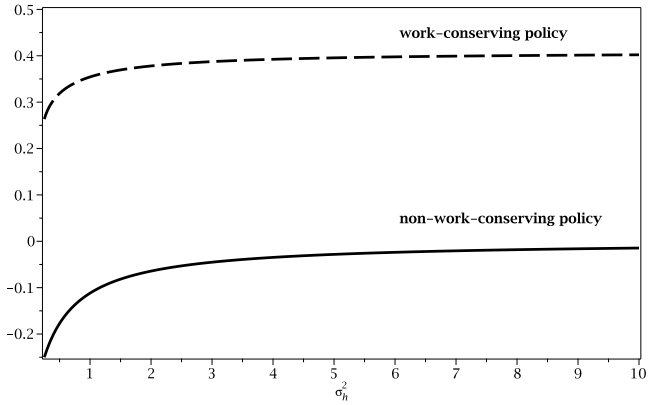


Figure 5.3: Correlation coefficient between the system contents for $\lambda \uparrow \frac{1}{2}$ versus the asymptotic arrival variance σ_h^2.

5.5 Concluding remarks

We compared two similar, albeit slightly different, discrete-time two-class queueing models that fall within a class of queueing models that are known to be hard to analyze. In this chapter, we have combined two well-known methods, namely the heavy-traffic limit method and the boundary-value method. By combining these two approaches, we have succeeded in deriving easy-to-use formulas for the correlation coefficient between the numbers of type-1 and type-2 customers, when the queues are close to saturation. We emphasize that we have assumed a symmetrical arrival process, but that it can have any distribution. The results reveal that the two queueing models are in fact very different from each other in heavy-traffic, which goes against our prior intuition.

6

Conclusions

In this last chapter, we end our journey and summarize the contributions of this dissertation and describe some interesting future directions.

6.1 Overview of the main contributions

In this dissertation, we have studied a two-class queueing system with a randomly alternating service discipline. The numbers of arrivals in the two queues are assumed to be general but independent from slot to slot and the service times of the customers are assumed to be deterministically equal to a single slot. In each time slot, the single server is available to queue-1 (queue-2) customers with probability α (resp. $1-\alpha$). If a queue happens to be empty at the moment that the server is allocated to that queue, no-one gets service during that slot.

In Chapter 2, various special cases from the perspective of the nature of the joint arrival distribution are analyzed in detail. The first case considers independent Bernoulli arrivals in the two queues. The second case considers identical Bernoulli arrivals in the two queues. The third case considers one stream of geometric arrivals and probabilistic routing to the two queues. The main contributions of this chapter are listed below.

- We succeeded in obtaining exact, closed-form, expressions for the joint probability distribution of the numbers of customers in both queues.
- The work in Chapter 2 laid foundations for the work done in Chapter 3.
- Exact, closed-form, expressions for several interesting numerical characteristics were provided.
- By means of numerical examples, we illustrated the influence of the service discipline on some of the most important performance measures of the queueing model.

- We determined a large class of arrival distributions such that the numbers of type-1 and type-2 customers in the system are independent.

We performed an asymptotic queueing analysis in Chapter 3. The results are related to the asymptotic analysis of random walks in the quarter plane. Hereafter, the main contributions of this chapter are summarized.

- We obtained an intriguing condition for the dominant singularities of the functions $U(z,0)$ and $U(0,z)$ to be simple poles.
- Asymptotic expressions for the joint probabilities of the numbers of customers in the system are obtained.
- In particular, we showed that it is possible to obtain asymptotic expressions for the stationary distribution of a random walk in the quarter plane for which the one-step displacements are not restricted to neighboring states.
- In the case of independent arrivals at the two queues, additional results were obtained. More concretely, in this case we showed that it is possible to determine the second most dominant singularity of $U(z,0)$ and $U(0,z)$ in case of high loads.

The asymptotic results of Chapter 3 do not give rise to accurate results for probabilities that are not in the tail. An approximation method for the latter probabilities is the subject of Chapter 4. The main contributions are listed here.

- We developed a novel approximation method.
- We showed that our results are highly accurate in case of light to moderate loads.
- We discussed the possible problems in case of high loads and gave some examples to illustrate our results.

In Chapter 5, we investigated the joint system-content distribution in heavy-traffic. To that end, we restricted ourselves to a symmetric system. A functional equation for the joint LST of the scaled system contents was derived. The main results we obtained are the following.

- The joint LST was obtained explicitly.
- We obtained a closed-form expression for the correlation coefficient between the numbers of customers in both queues when the queueing system is brought to the border of instability.
- We applied the same methodology to a slightly modified queueing model and compared the correlation coefficients of both models in heavy-traffic.

6.2 Future research

The *to-do list* with possible interesting extensions has kept growing during the previous four years. To conclude this dissertation, we describe some of the bullet points of this to-do list.

In this dissertation, we have assumed that the state of the server (available to either queue-1 or queue-2) changes independently from slot to slot. This feature was modeled by means of a sequence of i.i.d. random variables with common Bernoulli distribution. A logical extension of our queueing model, is to **model the state of the server by a two-state Markovian process**. This means that the state of the server in a slot depends on the state of the server during the previous slot. This server process is then specified by means of two parameters, for instance the conditional probabilities that the server is allocated to the first queue or the second queue during two consecutive time slots.

Another possible direction for future research could be to include **switch-over times**, see for example [135] for a closely related model and [136] for a survey paper. The switch-over time is defined as the number of slots taken for the server to switch to the other queue. Note that no customers can be served during switch-over times. If we want to include switch-over times, we have to take into account the state of the server. If the server switches, the number of arrivals during the switch-over time should be included in the system equations. The switch-over times can be modeled by i.i.d. random variables with a common PGF. To simplify the analysis, we would assume that the arrivals are independent of the switch-over times. Although the system equations become more complicated, we believe that an analysis as per Chapter 3 and/or Chapter 4 is possible. Finally, we remark that in case of (on average) long switch-over times, our non-work conserving queueing model might outperform its work-conserving variant. This is because in the case of long switch-over times, it might be sometimes beneficial to stay at an empty queue to anticipate for arrivals in the subsequent slots.

We have assumed a symmetrical system in Chapter 5 to establish the **heavy-traffic limit**. It would be interesting to obtain such a limit in the case of a non-symmetric system. However, in the **non-symmetric** model it is not obvious how to scale the system contents because our model has two stability conditions. Mathematically, we have to take the limits $\lambda_1 \to \alpha$ and $\lambda_2 \to 1-\alpha$ simultaneously. One possible approach we can think of to accomplish this, is to keep the ratio $\frac{\lambda_1}{\lambda_2}$ constant. Furthermore, we note that the computations of the boundary-value problem in the non-symmetric case might be trickier in comparison with the symmetric case. Another research direction is to incorporate correlation between the type-1 and type-2 arrivals. For example, one could consider a joint arrival PGF of the form $A(z_1, z_2) = C(z_1 z_2)$ or $A(z_1, z_2) = C(\frac{z_1}{2} + \frac{z_2}{2})$, with $C(z)$ a one-dimensional PGF. If $\alpha = \frac{1}{2}$, these two PGFs lead to a symmetric system yet again. Therefore, we believe that a similar

analysis as in Chapter 5 is feasible. Note that $A(z_1, z_2) = C(z_1 z_2)$ corresponds to the case of identical arrivals in the two queues and $A(z_1, z_2) = C(\frac{z_1}{2} + \frac{z_2}{2})$ corresponds to arrivals that are routed to the queues with equal probability.

Closely related to the concept of heavy-traffic approximation is that of **light-traffic approximation**. The latter approximation is usually easier to obtain than the heavy-traffic approximation. This is also true for the model studied in this dissertation, at least in the symmetric case. As in Chapter 5, let us assume a symmetric system such that $\lambda \triangleq \lambda_1 = \lambda_2$ and $\alpha = \frac{1}{2}$. If we expand $U(z_1, z_2)$ in a Taylor series in the variable λ, substitute this expansion into the fundamental functional equation and equate coefficients of corresponding powers of λ, we obtain for each coefficient a functional equation like (2.1). The upshot is that the kernel of this functional equation is equal to $z_1 z_2 - \frac{1}{2}(z_1 + z_2)$. This simple form of the kernel allows to obtain an explicit expression for the boundary functions present in the functional equation. We have found that the first four coefficients already provide a good approximation for small λ. A definite advantage of this approach in comparison with (for example) the approach of Chapter 4 is that a closed-form expression is obtained. Just like with the heavy-traffic approximation, it is not clear how to deal with an asymmetric system.

Finally, the most challenging future direction for the queueing model studied in this dissertation is the **generalization to higher-dimensional queueing models**. Although we emphasize that any knowledge gain for 3-dimensional problems is a huge leap forwards. This is because the generalization of the boundary-value approach to three dimensions is still an open problem. In the review paper [121] about the boundary-value method in queueing theory, the author indicates that the mathematical analysis as well as the numerical analysis are very complicated, based on the few attempts into this research direction. From an application point of view, considering 3 dimensions offers the possibility to investigate more complex queueing systems. From a mathematical point of view, there is a big gap in knowledge between 2-dimensional and 3-dimensional models.

Bibliography

[1] S. Asmussen, Applied Probability and Queues, Vol. 51, Springer Science & Business Media, 2008.

[2] H. Kobayashi, B. L. Mark, W. Turin, Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance, Cambridge University Press, 2011.

[3] M. Harchol-Balter, Performance Modeling and Design of Computer Systems: Queueing Theory in Action, Cambridge University Press, 2013.

[4] M. F. Neuts, Matrix-analytic methods in queuing theory, European Journal of Operational Research 15 (1) (1984) 2–12.

[5] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, 1999.

[6] D. Bini, G. Latouche, B. Meini, Numerical Methods for Structured Markov Chains, Oxford University Press, 2005.

[7] Z. Nehari, Conformal Mapping, Dover Publications, 1952.

[8] W. Feller, An Introduction to Probability Theory and its Applications, 3rd Edition, Vol. I, John Wiley & Sons, New York, 1957.

[9] L. Ahlfors, Complex Analysis, 3rd Edition, McGraw-Hill Education, 1979.

[10] M. J. Ablowitz, A. S. Fokas, Complex Variables: Introduction and Applications, Cambridge University Press, 2003.

[11] P. Flajolet, R. Sedgewick, Analytic Combinatorics, Cambridge University press, 2009.

[12] T. Takine, Queue length distribution in a FIFO single-server queue with multiple arrival streams having different service time distributions, Queueing Systems 39 (4) (2001) 349–375.

[13] H. Masuyama, T. Takine, Analysis and computation of the joint queue length distribution in a FIFO single-server queue with multiple batch Markovian arrival streams, Stochastic Models 19 (3) (2003) 149–381.

[14] S. De Clercq, K. Laevens, B. Steyaert, H. Bruneel, A multi-class discrete-time queueing system under the FCFS service discipline, Annals of Operations Research 202 (1) (2013) 59–73.

[15] J. Baetens, B. Steyaert, D. Claeys, H. Bruneel, System occupancy in a multiclass batch-service queueing system with limited variable service capacity, Annals of Operations Research 293 (1) (2020) 3–26.

[16] H. Takagi, Queueing Analysis: a Foundation of Performance Evaluation Volume 1: Vacation and Priority Systems, part 1, North-Holland, 1991.

[17] J. Walraevens, , B. Steyaert, H. Bruneel, Performance analysis of a single-server ATM queue with a priority scheduling, Computers & Operations Research 30 (12) (2003) 1807–1829.

[18] J. Walraevens, J. S. H. Leeuwaarden, O. J. Boxma, Power series approximations for generalized processor sharing systems, Queueing Systems 66 (2010) 107–130.

[19] J. Vanlerberghe, Analysis and optimization of discrete-time generalized processor sharing queues, Ph.D. thesis, Ghent University (2018).

[20] A. K. Parekh, R. G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: the single-node case, IEEE/ACM Transactions on Networking 1 (3) (1993) 344–357.

[21] H. Takagi, Queuing analysis of polling models, ACM Computing Surveys (CSUR) 20 (1) (1988) 5–28.

[22] H. Levy, M. Sidi, Polling systems: applications, modeling, and optimization, IEEE Transactions on Communications 38 (10) (1990) 1750–1760.

[23] V. M. Vishnevskii, O. V. Semenova, Mathematical methods to study the polling systems, Automation and Remote Control 67 (2) (2006) 173–220.

[24] S. Borst, O. Boxma, Polling: past, present, and perspective, Top 26 (3) (2018) 335–369.

[25] V. Vishnevsky, O. Semenova, Polling systems and their application to telecommunication networks, Mathematics 9 (2) (2021) 117.

[26] C. Dou, J. Chang, Serving two correlated queues with a synchronous server under exhaustive service discipline and nonzero switchover time, IEEE Transactions on Communications 39 (11) (1991) 1582–1589.

[27] H. Levy, L. Kleinrock, Polling systems with zero switch-over periods: a general method for analyzing the expected delay, Performance Evaluation 13 (2) (1991) 97–107.

[28] M. Eisenberg, Two queues with alternating service, SIAM Journal on Applied Mathematics 36 (2) (1979) 287–303.

[29] O. J. Boxma, W. P. Groenendijk, Two queues with alternating service and switching times, in: O. J. Boxma, R. Syski (Eds.), Queueing Theory and its Applications, Liber Amicorum for J.W. Cohen, North-Holland, Amsterdam, 1988, pp. 261–282.

[30] E. G. Coffman, G. Fayolle, I. Mitrani, Two queues with alternating service periods, in: Proceedings of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation, 1987, pp. 227–239.

[31] H. Takagi, K. Leung, Analysis of a discrete-time queueing system with time-limited service, Queueing Systems 18 (1) (1994) 183–197.

[32] K. Leung, Cyclic-service systems with nonpreemptive, time-limited service, IEEE Transactions on Communications 42 (8) (1994) 2521–2524.

[33] K. Leung, M. Eisenberg, A single-server queue with vacations and gated time-limited service, IEEE Transactions on Communications 38 (9) (1990) 1454–1462.

[34] K. Leung, M. Eisenberg, A single-server queue with vacations and nongated time-limited service, Performance Evaluation 12 (2) (1991) 115–125.

[35] J. Xie, M. J. Fischer, C. M. Harris, Workload and waiting time in a fixed-time loop system, Computers & Operations Research 24 (8) (1997) 789–803.

[36] R. de Haan, R. J. Boucherie, J. van Ommeren, A polling model with an autonomous server, Queueing Systems 62 (3) (2009) 279–308.

[37] A. Al Hanbali, R. de Haan, R. J. Boucherie, J. van Ommeren, Time-limited polling systems with batch arrivals and phase-type service times, Annals of Operations Research 198 (1) (2012) 57–82.

[38] M. Saxena, O. Boxma, S. Kapodistria, R. Queija, Two queues with random time-limited polling, Probability and Mathematical Statistics 37 (2) (2017) 257–289.

[39] N. Dvir, R. Hassin, U. Yechiali, Strategic behaviour in a tandem queue with alternating server, Queueing Systems 96 (3) (2020) 205–244.

[40] H. Bruneel, B. G. Kim, Discrete-Time Models for Communication Systems Including ATM, Kluwer Academic Publisher, Boston, 1993.

[41] S. Wittevrongel, H. Bruneel, Tail distribution of the buffer occupancy in a discrete-time queue fed by general on/off sources, COST257 TD (97) 45.

[42] S. Wittevrongel, Discrete-time buffers with variable-length train arrivals, Electronics Letters 34 (18) (1998) 1719–1721.

[43] K. De Turck, D. Fiems, S. Wittevrongel, H. Bruneel, A Taylor series expansions approach to queues with train arrivals, in: 5th International ICST Conference on Performance Evaluation Methodologies and Tools (Valuetools), Paris, 2011, pp. 447–455.

[44] B. Feyaerts, S. De Vuyst, H. Bruneel, S. Wittevrongel, Analysis of discrete-time buffers with heterogeneous session-based arrivals and general session lengths, Computers & Operations Research 39 (12) (2012) 2905–2914.

[45] D. Fiems, T. Maertens, H. Bruneel, Queueing systems with different types of server interruptions, European Journal of Operational Research 188 (3) (2008) 838–845.

[46] R. O. LaMaire, An M/G/1 vacation model of an FDDI station, IEEE Journal on Selected Areas in Communications 9 (2) (1991) 257–264.

[47] R. Núnez-Queija, Sojourn times in a processor sharing queue with service interruptions, Queueing systems 34 (1) (2000) 351–386.

[48] J. Vanlerberghe, T. Maertens, J. Walraevens, S. De Vuyst, H. Bruneel, On the optimization of two-class work-conserving parameterized scheduling policies, 4OR 14 (3) (2016) 281–308.

[49] J. W. Cohen, O. J. Boxma, Boundary Value Problems in Queueing System Analysis, North-Holland, Amsterdam, 1983.

[50] A. Devos, J. Walraevens, T. Phung-Duc, H. Bruneel, Analysis of the queue lengths in a priority retrial queue with constant retrial policy, Journal of Industrial & Management Optimization 16 (6) (2020) 2813–2842.

[51] A. Devos, J. Walraevens, H. Bruneel, A priority retrial queue with constant retrial policy, in: International Conference on Queueing Theory and Network Applications, Springer, 2018, pp. 3–21.

[52] A. Devos, J. Walraevens, D. Fiems, H. Bruneel, Analysis of a discrete-time two-class randomly alternating service model with Bernoulli arrivals, Queueing Systems 96 (1-2, SI) (2020) 133–152.

[53] G. Fayolle, R. Iasnogorodski, Two coupled processors: the reduction to a Riemann-Hilbert problem, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47 (3) (1979) 325–351.

[54] J. P. C. Blanc, Application of the theory of boundary value problems in the analysis of a queueing model with paired services., Ph.D. thesis, University of Utrecht (1982).

[55] M. Noufissa, A two-node Jackson's network subject to breakdowns, Communications in Statistics. Stochastic Models 4 (3) (1988) 523–552.

[56] W. Feng, M. Kowada, K. Adachi, A two-queue model with Bernoulli service schedule and switching times, Queueing Systems 30 (3) (1998) 405–434.

[57] G. Fayolle, V. A. Malyshev, R. Iasnogorodski, Random Walks in the Quarter-plane, Vol. 40, Springer, 1999.

[58] P. Nain, G. Vardoyan, S. Guha, D. Towsley, Analysis of a tripartite entanglement distribution switch, Submitted, hal:03195985v4 (2021).

[59] K. Avrachenkov, P. Nain, U. Yechiali, A retrial system with two input streams and two orbit queues, Queueing Systems 77 (1) (2014) 1–31.

[60] S. Kapodistria, Z. Palmowski, Matrix geometric approach for random walks: Stability condition and equilibrium distribution, Stochastic Models 33 (4) (2017) 572–597.

[61] I. Dimitriou, A queueing model with two classes of retrial customers and paired services, Annals of Operations Research 238 (1-2) (2016) 123–143.

[62] I. Dimitriou, A two-class retrial system with coupled orbit queues, Probability in the Engineering and Informational Sciences 31 (2) (2017) 139–179.

[63] I. Dimitriou, N. Pappas, Performance analysis of an adaptive queue-aware random access scheme with random traffic, in: 2018 IEEE International Conference on Communications (ICC), 2018, pp. 1–6.

[64] I. Dimitriou, N. Pappas, Stable throughput and delay analysis of a random access network with queue-aware transmission, IEEE Transactions on Wireless Communications 17 (5) (2018) 3170–3184.

[65] I. Dimitriou, N. Pappas, Performance analysis of a cooperative wireless network with adaptive relays, Ad Hoc Networks 87 (2019) 157–173.

[66] I. Dimitriou, Analysis of the symmetric join the shortest orbit queue, Operations Research Letters 49 (1) (2021) 23–29.

[67] F. Guillemin, C. Knessl, J. S. H. van Leeuwaarden, Wireless three-hop networks with stealing II: exact solutions through boundary value problems, Queueing Systems 74 (2) (2013) 235–272.

[68] C. Fricker, F. Guillemin, P. Robert, G. Thompson, Analysis of an offloading scheme for data centers in the framework of fog computing, ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS) 1 (4) (2016) 1–18.

[69] J. S. H. Van Leeuwaarden, J. A. C. Resing, A tandem queue with coupled processors: computational issues, Queueing Systems 51 (1) (2005) 29–52.

[70] R. L. Disney, D. König, Queueing networks: A survey of their random processes, SIAM Review 27 (3) (1985) 335–403.

[71] J. Kingman, Two similar queues in parallel, The Annals of Mathematical Statistics 32 (4) (1961) 1314–1323.

[72] L. Flatto, H. P. McKean, Two queues in parallel, Communications on Pure and Applied Mathematics 30 (2) (1977) 255–263.

[73] S. Jaffe, The equilibrium distribution for a clocked buffered switch, Probability in the Engineering and Informational Sciences 6 (4) (1992) 425–438.

[74] J. W. Cohen, On the analysis of the symmetrical shortest queue, Tech. Rep. BS-R9420, Department of Operations Research, Statistics, and System Theory, CWI, Amsterdam (1994).

[75] J. W. Cohen, On the determination of the stationary distribution of a symmetric clocked buffered switch, in: Teletraffic Science and Engineering, Vol. 2, Elsevier, 1997, pp. 297–307.

[76] J. W. Cohen, On the asymmetric clocked buffered switch, Queueing Systems 30 (3) (1998) 385–404.

[77] J. W. Cohen, Analysis of the asymmetrical shortest two-server queueing model, Journal of Applied Mathematics and Stochastic Analysis 11 (2) (1998) 115–162.

[78] J. W. Cohen, On a class of two-dimensional nearest-neighbour random walks, Journal of Applied Probability 31 (A) (1994) 207–237.

[79] L. Flatto, S. Hahn, Two parallel queues created by arrivals with two demands I, SIAM Journal on Applied Mathematics 44 (5) (1984) 1041–1053.

[80] P. E. Wright, Two parallel processors with coupled inputs, Advances in Applied Probability 24 (4) (1992) 986–1007.

[81] A. G. Konheim, I. Meilijson, A. Melkman, Processor-sharing of two parallel lines, Journal of Applied Probability 18 (4) (1981) 952–956.

[82] L. Haque, Y. Q. Zhao, L. Liu, Sufficient conditions for a geometric tail in a QBD process with many countable levels and phases, Stochastic Models 21 (1) (2005) 77–99.

[83] Q. M. He, H. Li, Y. Q. Zhao, Light-tailed behavior in QBD processes with countably many phases, Stochastic Models 25 (1) (2009) 50–75.

[84] M. Miyazawa, Tail decay rates in double QBD processes and related reflected random walks, Mathematics of Operations Research 34 (3) (2009) 547–575.

[85] M. Kobayashi, M. Miyazawa, Tail asymptotics of the stationary distribution of a two-dimensional reflecting random walk with unbounded upward jumps, Advances in Applied Probability 46 (2) (2014) 365–399.

[86] H. Li, Y. Q. Zhao, Tail asymptotics for a generalized two-demand queueing model – a kernel method, Queueing Systems 69 (1) (2011) 77–100.

[87] H. Li, Y. Zhao, A kernel method for exact tail asymptotics - random walks in the quarter plane, Queueing Models and Service Management 1 (1) (2018) 95–129.

[88] M. Kobayashi, M. Miyazawa, Revisiting the tail asymptotics of the double QBD process: refinement and complete solutions for the coordinate and diagonal directions, in: Matrix-analytic Methods in Stochastic Models, Springer, 2013, pp. 145–185.

[89] F. Guillemin, J. S. H. van Leeuwaarden, Rare event asymptotics for a random walk in the quarter plane, Queueing Systems 67 (1) (2011) 1–32.

[90] T. Ozawa, Asymptotics for the stationary distribution in a discrete-time two-dimensional quasi-birth-and-death process, Queueing Systems 74 (2) (2013) 109–149.

[91] T. Ozawa, M. Kobayashi, Exact asymptotic formulae of the stationary distribution of a discrete-time two-dimensional QBD process, Queueing Systems 90 (3) (2018) 351–403.

[92] A. Shwartz, A. Weiss, Large Deviations for Performance Analysis: Queues, Communication and Computing, Vol. 5, CRC Press, 1995.

[93] F. Den Hollander, Large Deviations, Vol. 14, American Mathematical Society, 2008.

[94] J. Walraevens, D. Claeys, T. Phung-Duc, Asymptotics of queue length distributions in priority retrial queues, Performance Evaluation 127 (2018) 235–252.

[95] H. S. Wilf, Generatingfunctionology, Academic press, 1990.

[96] M. Gonzalez, Classical Complex Analysis, CRC Press, Boca Raton, FL, USA, 1991.

[97] P. Kravanja, M. Van Barel, O. Ragos, M. N. Vrahatis, F. A. Zafiropoulos, Zeal: A mathematical software package for computing zeros of analytic functions, Computer Physics Communications 124 (2-3) (2000) 212–232.

[98] A. Devos, J. Walraevens, D. Fiems, H. Bruneel, Approximations for the performance evaluation of a discrete-time two-class queue with an alternating service discipline, Annals of Operations Research (2020) 1–27.

[99] J. Vanlerberghe, J. Walraevens, T. Maertens, H. Bruneel, A procedure to approximate the mean queue content in a discrete-time generalized processor sharing queue with Bernoulli arrivals, Performance Evaluation 134 (2019) 102001.

[100] J. P. C. Blanc, Performance analysis and optimization with the power-series algorithm, in: Performance Evaluation of Computer and Communication Systems, Springer, 1993, pp. 53–80.

[101] G. Hooghiemstra, M. Keane, S. Van de Ree, Power series for stationary distributions of coupled processor models, SIAM Journal on Applied Mathematics 48 (5) (1988) 1159–1166.

[102] I. J. Adan, O. J. Boxma, J. A. C. Resing, Queueing models with multiple waiting lines, Queueing Systems 37 (1) (2001) 65–98.

[103] D. Fiems, T. Phung-Duc, Light-traffic analysis of random access systems without collisions, Annals of Operations Research 277 (2) (2019) 311–327.

[104] E. Altman, K. E. Avrachenkov, R. Núnez-Queija, Perturbation analysis for denumerable Markov chains with application to queueing models, Advances in Applied Probability 36 (3) (2004) 839–853.

[105] M. Saxena, S. Kapodistria, R. Núnez-Queija, Perturbation analysis of two queues with random time-limited polling, in: the short paper conference proceedings of the 14th International Conference on Queueing Theory and Network Applications (QTNA2019), 2019.

[106] I. J. Adan, J. Wessels, W. H. M. Zijm, A compensation approach for two-dimensional Markov processes, Advances in Applied Probability 25 (4) (1993) 783–817.

[107] I. J. Adan, S. Kapodistria, J. S. van Leeuwaarden, Erlang arrivals joining the shorter queue, Queueing Systems 74 (2) (2013) 273–302.

[108] J. Selen, I. J. Adan, S. Kapodistria, J. van Leeuwaarden, Steady-state analysis of shortest expected delay routing, Queueing Systems 84 (3) (2016) 309–354.

[109] I. Dimitriou, Analysis of the symmetric join the shortest orbit queue, Operations Research Letters 49 (1) (2021) 23–29.

[110] G. J. van Houtum, I. J. Adan, J. Wessels, W. H. M. Zijm, The compensation approach for three or more dimensional random walks, in: DGOR/ÖGOR, Springer, 1993, pp. 342–349.

[111] I. J. Adan, O. J. Boxma, S. Kapodistria, V. G. Kulkarni, The shorter queue polling model, Annals of Operations Research 241 (1) (2016) 167–200.

[112] I. J. Adan, J. Wessels, Shortest expected delay routing for Erlang servers, Queueing systems 23 (1) (1996) 77–105.

[113] M. Saxena, I. Dimitriou, S. Kapodistria, Analysis of the shortest relay queue policy in a cooperative random access network with collisions, Queueing Systems 94 (1) (2020) 39–75.

[114] Y. Sakuma, M. Miyazawa, On the effect of finite buffer truncation in a two-node Jackson network, Journal of Applied Probability 42 (1) (2005) 199–222.

[115] J. Van Velthoven, B. Van Houdt, C. Blondia, The impact of buffer finiteness on the loss rate in a priority queueing system, in: European Performance Engineering Workshop, Springer, 2006, pp. 211–225.

[116] G. Latouche, G. T. Nguyen, P. G. Taylor, Queues with boundary assistance: the effects of truncation, Queueing systems 69 (2) (2011) 175–197.

[117] Y. Takahashi, K. Fujimoto, N. Makimoto, Geometric decay of the steady-state probabilities in a quasi-birth-and-death process with a countable number of phases, Stochastic Models 17 (1) (2001) 1–24.

[118] N. Van Dijk, Error bounds and comparison results: the Markov reward approach for queueing networks, in: Queueing Networks, Springer, 2011, pp. 397–459.

[119] Y. Chen, Random walks in the quarter-plane: Invariant measures and performance bounds, Ph.D. thesis, University of Twente (2015).

[120] X. Bai, Performance bounds for random walks in the positive orthant, Ph.D. thesis, University of Twente (2018).

[121] J. W. Cohen, Boundary value problems in queueing theory, Queueing Systems 3 (2) (1988) 97–128.

[122] H. C. Tijms, D. J. Van Vuuren, Markov processes on a semi-infinite strip and the geometric tail algorithm, Annals of Operations Research 113 (1) (2002) 133–140.

[123] R. Timmerman, Performance analysis at the crossroad of queueing theory and road traffic, Ph.D. thesis, Eindhoven University of Technology (2022).

[124] J. F. C. Kingman, The single server queue in heavy traffic, in: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 57, Cambridge University Press, 1961, pp. 902–904.

[125] C. Knessl, On the diffusion approximation to two parallel queues with processor sharing, IEEE Transactions on Automatic Control 36 (12) (1991) 1356–1367.

[126] J. A. Morrison, Diffusion approximation for head-of-the-line processor sharing for two parallel queues, SIAM Journal on Applied Mathematics 53 (2) (1993) 471–490.

[127] C. Knessl, J. A. Morrison, Heavy traffic analysis of two coupled processors, Queueing Systems 43 (3) (2003) 173–220.

[128] G. Foschini, J. Salz, A basic dynamic routing problem and diffusion, IEEE Transactions on Communications 26 (3) (1978) 320–327.

[129] F. I. Karpelevich, A. Y. Kreinin, Two-phase queuing system GI/G/1→ G'/1/∞ under heavy traffic conditions, Theory of Probability & Its Applications 26 (2) (1982) 293–313.

[130] M. Saxena, Two-dimensional queueing models, Ph.D. thesis, Eindhoven University of Technology (2020).

[131] S. Kapodistria, M. Saxena, O. Boxma, O. Kella, Workload analysis of a two-queue fluid polling model, Submitted, arXiv preprint arXiv:2112.04819.

[132] A. Devos, J. Walraevens, D. Fiems, H. Bruneel, Heavy-traffic comparison of a discrete-time generalized processor sharing queue and a pure randomly alternating service queue, Mathematics 9 (21) (2021) 2723.

[133] M. A. Lawrentjew, B. V. Shabat, Methoden der Komplexen Funktionentheorie, VEB Deutscher Verlag der Wissenschaften, Berlin, Germany, 1967.

[134] R. Kress, Numerical Analysis, New York (N.Y.) : Springer, 1998.

[135] G. D. Celik, L. B. Le, E. Modiano, Scheduling in parallel queues with randomly varying connectivity and switchover delay, in: 2011 Proceedings IEEE INFOCOM, 2011, pp. 316–320.

[136] K. Aziz, Effect of switchover time in cyclically switched systems, in: J. Kleban (Ed.), Switched Systems, IntechOpen, Rijeka, 2009, Ch. 6.

A queue of people, waiting to be served.