Faculteit Toegepaste Wetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter : Prof. Dr. ir. Herwig Bruneel

Analyse van discrete-tijd-wachtlijnsystemen met vakanties

Analysis of discrete-time queueing systems with vacations

door

ing. Dieter Fiems

Promotor : Prof. Dr. ir. Herwig Bruneel

PROEFSCHRIFT INGEDIEND BIJ DE FACULTEIT TOEGEPASTE WETENSCHAPPEN VAN DE UNIVERSITEIT GENT TOT HET BEHALEN VAN DE ACADEMISCHE GRAAD VAN DOCTOR IN DE TOEGEPASTE WETENSCHAPPEN

Academiejaar 2003–2004

Acknowledgements

You are about to start reading the transcript of my research at the Department of Telecommunications and Information Processing of Ghent University. Although it seems like yesterday, it is more than 5 years ago that I started working on queueing theory. Pleasant years that have passed by quickly.

There are many persons that I have to thank, who have directly or indirectly contributed to this dissertation and without whom life at work and out of work would have been a lot less enjoyable.

First of all, I would like to thank my advisor, Professor Herwig Bruneel. His initial guidance and everlasting encouragement and support helped me to complete this dissertation.

Further, there are the colleagues at the Department of Telecommunication and Information Processing, always ready to lend a helping hand. Bart Steyaert patiently taught me the art of scientific paper writing. My office companions Joris Walraevens and Stijn De Vuyst were always happy to engage into fruitful and enlightening discussions – ranging from “the hitchhikers guide” to “complex contour integration” – that often lead to new insights and ideas.

Finally, I would like to thank my family and friends for their continuing support.

Dieter Fiems

Contents

1 Introduction **1**

1.1 Background . . . 2

1.1.1 Applications of vacation models . . . 2

1.1.2 Discrete-time queueing theory . . . 5

1.1.3 The probability generating functions approach . . . 5

1.1.4 Kendall's notation . . . 6

1.2 The discrete-time $Geo^X/G/1$ queue . . . 7

1.2.1 System description and notation . . . 7

1.2.2 Analysis at departure epochs . . . 8

1.2.3 Analysis at random slot boundaries . . . 15

1.2.4 Analysis at arrival epochs . . . 19

1.2.5 Little's result . . . 20

1.3 Overview . . . 23

2 Random server vacations **25**

2.1 Bernoulli vacations . . . 27

2.1.1 Queueing model . . . 27

2.1.2 Effective service times . . . 28

2.1.3 Queue content and customer delay . . . 36

2.1.4 Some numerical results . . . 37

2.2 The method of the supplementary variables . . . 38

2.2.1 Continue after interruption . . . 40

2.2.2 Repeat after interruption with resampling . . . 43

2.2.3 Repeat after interruption . . . 46

2.2.4 Performance measures . . . 51

2.2.5 Comparison . . . 52

2.3 Two-state Markovian vacations . . . 53

2.3.1 Vacation process . . . 54

2.3.2 Effective service times . . . 55

2.3.3 Queue content . . . 62

2.3.4 Unfinished Work and Customer Delay . . . 67

2.3.5 Numerical example . . . 70

2.4 Generally distributed vacations . . . 74

2.4.1 Vacation process 75
2.4.2 Effective service times 75
2.4.3 Queue content and customer delay 82
2.4.4 Idle and busy periods 88
2.4.5 Numerical example 90
2.5 Preemptive priority queueing systems 92
2.5.1 Priority models 93
2.5.2 Queueing model and analysis 94
2.5.3 Approximate analysis 95
2.5.4 Numerical examples 97
2.6 Other operation modes 101
2.6.1 Delayed operation modes 102
2.6.2 Partial interruption modes 105
2.6.3 Some numerical examples 107

3 Other vacation queues 113
3.1 The gated-exhaustive vacation queue 116
3.1.1 Mathematical model 116
3.1.2 System equations 117
3.1.3 The joint probability generating functions 120
3.1.4 Queue content at various epochs 124
3.1.5 Customer delay 126
3.1.6 Special cases 133
3.1.7 Numerical Examples 135
3.2 Non-gated vacation queues 140
3.2.1 Mathematical Model 140
3.2.2 Service completion times 141
3.2.3 Queue content 146
3.2.4 Special cases I: systems without service interruptions 148
3.2.5 Special cases II: systems with service interruptions 151
3.2.6 Numerical example 153

4 Summary 155
4.1 Random service vacations 155
4.2 Other vacation models 156

A Frequently used distributions 159

B From discrete to continuous time 163
B.1 The continuous-time model 163
B.2 Adapting arrival and service processes 164
B.3 Taking the limit 165

Samenvatting **169**
S.1 Inleiding 169
S.1.1 Toepassingen van wachtlijnsystemen met vakanties 170
S.1.2 Discrete-tijd-wachtlijnsystemen 173
S.1.3 De methode van de probabiliteitsgenererende functies 174
S.1.4 Overzicht 175
S.2 Willekeurige vakanties 176
S.2.1 Algemene onderstellingen 176
S.2.2 Vakantiemodellen 177
S.2.3 Onderbroken bediening 179
S.2.4 De methode van de effectieve bedieningstijden 181
S.2.5 Toepassing: een preëmptief prioriteitsmodel 185
S.3 Andere vakantiemodellen 189
S.3.1 Klassieke vakantiemodellen 190
S.3.2 Een vakantiemodel met een poort en exhaustieve bediening . . 191
S.3.3 Een raamwerk voor vakantiemodellen zonder poorten 194

Notation

Regarding notation, one often has to choose between consistency and readability. We have tried to be as consistent as possible without sacrificing readability.

Systematics

X, Y	random variables
$x(i)$	the probability mass function of the random variable X
$X(z)$	the probability generating function of the random variable X
$p(i,j)$	the joint probability mass function of two random variables, the exact meaning is clarified in the context
$x(i\|j)$	the probability mass function of a random variable X conditioned on $Y = j$, the meaning of Y is clarified in the context
$P(x,y)$	the joint probability generating function of two random variables, the exact meaning is clarified in the context
$X(z\|j)$	the probability generating function of random variable X conditioned on $Y = j$, the meaning of Y is clarified in the context
μ_X	the mean value of the random variable X
${\sigma_X}^2$	the variance of the random variable X
γ_X	the skewness of the random variable X
$\hat{X}, \tilde{X}, \check{X} \ldots$	mathematical accents are used to differentiate between related random variables

System variables

σ	the mean server availability
K	the burstiness factor of the vacation process
ρ	the system load

Some operators

$\mathrm{Pr}[\,]$	the probability operator
$\mathrm{E}[\,]$	the mean value operator
$(\,\cdot\,)^+$	shorthand notation for $\max(\,\cdot\,,0)$

Random variables

Ω, Θ, ...	Greek symbols denote auxiliary random variables, their meaning is clarified in the context
A	the length of an available period
B	the length of a vacation
C	the service completion time of a customer
D	the delay of a customer
E	the number of customer arrivals during a slot
F	the remaining vacation time
G	the total service time of the customer in service at a slot boundary
H	the remaining service time at slot boundaries of the customer in service
I	the time between consecutive batches of arrivals
Q	the number of available servers during a slot
R	the total unfinished service time at slot boundaries
S	the service time of a customer
T	the effective service time of a customer
U	the unfinished work at slot boundaries
V_a	the queue content as seen by arriving customers
V_d	the queue content at customer departure epochs
V_r	the queue content at slot boundaries
V_s	the queue content when a customer starts service
V_v	the queue content at the beginning of a random vacation slot
W	the waiting time of a customer
X	the sub-busy period of a customer
Y_B	the length of a busy period
Y_I	the length of an idle period

Acronyms

CAI	continue after interruption
RAI	repeat after interruption
RAI,wr	repeat after interruption with resampling
d-RAI	delayed repeat after interruption
d-RAI,wr	delayed repeat after interruption with resampling
p-RAI	partial repeat after interruption
p-RAI,wr	partial repeat after interruption with resampling
dp-RAI	delayed partial repeat after interruption
dp-RAI,wr	delayed partial repeat after interruption with resampling
SV	supplementary variable
EST	effective service time
FIFO	first-in-first-out
i.i.d.	independent and identically distributed
ATM	asynchronous transfer mode
ARQ	automatic repeat request

Chapter 1

Introduction

Adde parvum parvo magnus acervus erit.

(Add little to little and there will be a big pile)

– Ovid

To some extent, waiting in line is part of everyday life. Whether one considers waiting lines at the checkout counter in a grocery store, traffic jams during rush hour or delays while browsing the world wide web, in essence the same phenomenon occurs. Several customers demand some kind of service at the same time, while this service is only available to a limited number of customers simultaneously. Therefore, some of these customers have to wait until they can be served, i.e., they have to queue.

Abstractly, a queueing system consists of *customers*, waiting in *queues* or *buffers* until they can be served. For a particular queue, there may be several parallel service facilities or *servers* and one refers to the amount of time a customer occupies a server as his *service time*. Often customers are served in order of their arrival but the server may select customers according to some other rule or *service discipline*. Once a customer has received service, he may either leave the queueing system or join some other (or even the same) queue.

Queueing theory investigates queueing phenomena in a stochastic framework. That is, stochastic processes capture the uncertainty regarding customer arrival times and their service requests. The queueing theoretician then relates stochastic properties of the arrival and service processes to performance measures of the queueing system such as moments of the *queue content* at various points in time and of the *customer delay*. The terms queue content and customer delay refer to the number of customers in the queueing system, including customers in service and to the time a customer spends in the queueing system respectively.

In this dissertation, we focus on a particular kind of queueing systems. All systems under consideration have a single server at their disposal. The latter however, is unavailable to waiting customers from time to time. In queueing parlance, one says that the server leaves for a *vacation* and these models are therefore referred to as vacation models, hence the title of this dissertation. Various vacation models are presented and analysed in chapters 2 and 3. We first situate our research topic in the next section. As an introduction to the probability generating functions approach which is used throughout this dissertation, we present the analysis of a queueing system without vacations in section 1.2. This will also allow us to settle some notation. We conclude this chapter with an outline of the following chapters.

1.1 Background

The core contribution of this dissertation are the analyses of various queueing systems with vacations. We consider queueing phenomena in a discrete-time framework. That is, we assume that time is divided into fixed length intervals or slots. Further, the analyses are based on frequent use of z-transforms of probability mass functions, the so-called probability generating functions. To situate our work, we therefore focus on the following three key aspects: queueing systems with vacations, the discrete-time assumption and the probability generating functions approach. For further use, we also introduce Kendall's notation for queueing systems.

1.1.1 Applications of vacation models

In the past, queueing systems with vacations have received considerable attention in literature. Doshi's survey paper [1986] as well as Takagi's monographs [1991, 1993] point to numerous contributions on vacation systems. We postpone a profound survey of vacation models from a mathematical point of view to the beginning of chapters 2 and 3. Here, we situate vacation models by reviewing some literature on applications of this type of queueing models. Vacation models turn out to be particularly useful in the performance assessment of multi-class queueing systems and of queueing systems with unreliable servers.

Multi-class systems

Queueing models with vacations can be used to assess the performance of systems where different classes of customers contend for access to the same server(s) such as priority systems or polling systems. Figure 1.1 depicts such a multi-class system. Customers only receive service when the switch is turned to their buffer. Therefore,

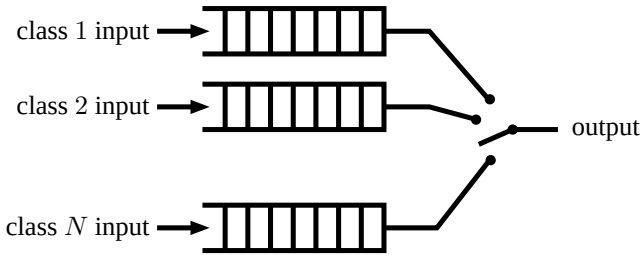


Figure 1.1: A multi-class queueing system.

customers of a particular class perceive an unavailable server when customers of different classes are served. That is, from the vantage point of the customers of that particular class, their server leaves for a vacation from time to time.

Vacation models are used by amongst others Avi-Itzhak and Naor [1963], Nain [1983] and Gaver Jr. [1962] to investigate priority queueing systems. The former contributions consider a preemptive resume priority discipline whereas the latter contribution considers the preemptive repeat and the preemptive repeat with resampling disciplines. Preemption means that service of lower class customers is interrupted when higher class customers arrive. Lower class customers then either resume service (preemptive resume) or repeat service with the same (preemptive repeat) or with a possibly different (preemptive repeat with resampling) service time when all higher class customers are served. For preemptive priority disciplines, service periods of higher class customers are perceived as server vacations by lower class customers. It can be shown that the performance of preemptive priority queueing systems can be assessed by means of queueing models with a *random vacation process*. This particular type of vacation models is investigated in chapter 2. In particular, section 2.5 addresses the performance assessment of preemptive priority queueing systems by means of vacation models.

Queueing systems with vacations are also used to model queueing systems with a priority scheduling discipline in specific contexts. Bruneel [1983a] investigates the performance of data traffic sent during the silent periods of a voice channel. In this case, voice traffic has priority over data traffic. Further, Núñez Queija [2000] investigates the performance of available bit rate (ABR) class traffic in an asynchronous transfer mode (ATM) network. This type of traffic can only be sent when there is no other traffic. Finally, Cowan [1987] uses a vacation model to assess the queueing delays of cars on an unsignalised intersection of a minor and major road. Cars on the major road have priority over cars on the minor road.

A polling system periodically checks the queues of the different classes of customers. If there are customers present, the server serves some or all of these. Of course, other class customers perceive a server vacation when the latter is serving customers of other classes. Queueing models with vacations for polling systems are suggested

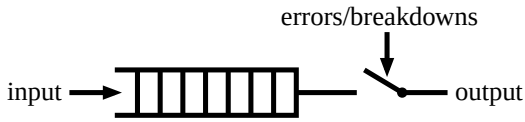


Figure 1.2: A queueing system with errors or breakdowns.

and analysed by a.o. LaMaire [1991], Leung and Lucantoni [1994] and Chiarawongse et al. [1994]. In these contributions, polling corresponds to token passing in token ring or token bus networks. The corresponding vacation models are number-limited [LaMaire, 1991] and time-limited [Leung and Lucantoni, 1994, Chiarawongse et al., 1994] multiple vacation models. We will explain and investigate these particular types of vacation models in more detail in chapter 3.

Unreliable servers

In many practical queueing situations, the server is unreliable from time to time. Unreliability either results in service errors or in service breakdowns. Service errors require that a customer receives service again whenever there were errors during his service. During service breakdowns on the other hand, customers are no longer served and some user action is required to reestablish proper server behaviour. That is, the server needs maintenance. Figure 1.2 depicts an abstract queueing system with errors or breakdowns. Customers are only served (properly) when the switch is closed.

An example of a system with an unreliable server can be found in wireless communication as transmission of data packets over a wireless communication channel is often error-prone. The communication channel is therefore perceived as one that takes vacations from time to time. Unfortunately, the transmitter cannot detect whether or not the channel is available. Therefore, the receiver sends feedback to acknowledge correct (or incorrect) reception of packets according to an automatic repeat request (ARQ) protocol such as the stop-and-wait, the go-back-N or the selective repeat protocol. Performance analyses of ARQ protocols yield some interesting queueing problems. Due to the limited capacity of the communication channel, queueing arises at the sender side. Further, as transmitted data may be received out of order, some data may have to wait (that is, to queue) for preceding data to arrive correctly on the receiver side. Towsley and Wolf [1979] and Towsley [1981] consider the stop-and-wait ARQ protocol whereas Towsley [1981] and Yoshimoto et al. [1993] consider the go-back-N ARQ protocol. For these protocols, packets are received in order and therefore queueing only arises at the sender side. This is not the case for the selective repeat protocol. Bruneel et al. [1990] and Shacham and Towsley [1991] focus on packet delay caused by resequencing at the receiver side, whereas Kim and Krunz [2000] present approximate results for the total experienced delay. The latter includes the queueing delay at the sender side and the resequencing delay at the receiver side.

Amongst others, Van der Duyn Schouten and Vanneste [1995], Wang et al. [1995] and Perry and Posner [2000] consider queueing systems with service breakdowns and preventive maintenance. Vacations here correspond to preventive maintenance jobs or to service breakdowns. Often, the time that the server needs to recover from a breakdown is much longer than the time spent on preventive maintenance. Therefore, system performance can improve by careful scheduling of preventive maintenance jobs.

1.1.2 Discrete-time queueing theory

In this dissertation, we always assume that time is divided into fixed length intervals or *slots*. That is, we assume a discrete-time scale. Customers arrive in the queueing system under consideration during the consecutive slots, but they can only start service at the beginning of slots. That is, service of customers is synchronised with respect to slot boundaries. Further, customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. One refers to this type of queueing systems as discrete-time queueing systems.

The discrete-time scale often reflects the nature of an underlying application: for example, the clock time unit in a computer system, fixed size data units (bits, bytes, fixed length packets) on a communication channel, etc .... For sufficiently small slot lengths, discrete-time queueing models may also be used as an approximation of corresponding models where the time scale is continuous. In fact, one can obtain results for continuous-time models directly from the equivalent discrete-time results. We illustrate this assertion in appendix B.

Discrete-time queueing systems have been a research topic for several decades now. Early investigations were made by amongst others Meisling [1958], Birdsall et al. [1962] and also by Powell and Avi-Itzhak [1967]. Reference works on discrete-time queueing theory include the monographs of Bruneel and Kim [1993], Woodward [1993] and Takagi [1993]. Further, Hunter [1983a,b] considers some discrete-time queueing models in his two-volume book on applied probability. More recently, Daduna [2001] considered networks of discrete-time queues.

1.1.3 The probability generating functions approach

Over the years, different methodologies have been developed to assess the performance of queueing systems. The two main analytical approaches are the matrix analytic method (see e.g., Neuts [1983]) and the transform method (see e.g., Bruneel and Kim [1993] and Takagi [1993] for discrete-time analyses and Kleinrock [1975, 1976] and Takagi [1991] for continuous-time analyses). The first method translates the queueing problem into a matrix equation. Computationally efficient algorithms are then developed to solve this equation, hereby exploiting the structural properties

of the involved matrix. The transform method relies on the use of generating functions – in particular Laplace-transforms and z-transforms – to facilitate the queueing analysis. Facilitation is brought about by the fact that the generating function of the sum of two independent random variables equals the product of the generating functions of these variables. The transition to the transform domain implies that one obtains generating functions (and not probability mass functions or density functions) of random variables under interest such as queue content in equilibrium or customer delay. The moment generating property of these transforms then allows one to obtain performance measures such as means and variances of these random variables, whereas a singularity analysis yields approximate expressions for tail probabilities. We will use the transform approach throughout this dissertation.

For increasingly complex queueing problems, analytical methods may fail. One therefore has to rely on simulation studies to assess the performance of such systems (see a.o., Law and Kelton [1991]). The key advantage of simulation is its generality. The methodology hardly puts limits on the complexity of the systems under consideration. This generality comes at the cost of time. That is, simulation can be very time consuming. This is especially the case if one is interested in the accurate estimation of the probability of events that do not occur frequently in time (rare events) or if one wants to assess the performance for numerous sets of parameters.

1.1.4 Kendall's notation

For further use, we here introduce Kendall's shorthand notation for queueing systems.

Kendall's shorthand notation for queueing systems allows one to specify the main characteristics of a queueing system concisely. A queueing system is characterised by the letter code $A/B/c$. Here A and B characterise the interarrival and the service times respectively whereas c corresponds to the number of servers.

Common continuous-time arrival processes include the Poisson arrival process ($A = M$), the Poisson batch arrival process ($A = M^X$), and processes with deterministic ($A = D$), Erlang distributed ($A = E_k$) and arbitrarily distributed ($A = G$) interarrival times. Service times may be a.o. deterministic ($B = D$), Erlang distributed ($B = E_k$) or arbitrarily distributed ($B = G$).

For discrete-time systems, interarrival times may be geometrically distributed ($A = Geo$), deterministic ($A = D$) or arbitrarily distributed ($A = G$). The discrete-time equivalent of the batch Poisson process is the discrete-time arrival process with geometrically distributed interarrival times of batches ($A = Geo^X$). Service times may be a.o. deterministic ($B = D$) or arbitrarily distributed ($B = G$).

Often, one sees from the type of arrival process whether the system under consideration is a discrete-time or a continuous-time system. However this is not always the

case. One may have to specify the time setting as well as particularities of the system such as server vacations, priority scheduling, etc, ...

1.2 The discrete-time $Geo^X/G/1$ queue

This section considers the discrete-time $Geo^X/G/1$ queueing system without vacations. This allows us to introduce general assumptions regarding discrete-time queueing systems, to introduce the probability generating functions approach and to fix some basic notation which will be used throughout the rest of this dissertation. Also, the results presented in section 2.1 depend on those presented here as we will see further.

Many authors have considered the $Geo^X/G/1$ queueing system before. Our references include amongst others Bruneel [1993], Bruneel and Kim [1993] and Takagi [1993]. The analysis as presented closely follows the lines of Takagi [1993], although we also include some joint probability generating function results as in Bruneel and Kim [1993].

1.2.1 System description and notation

We consider a discrete-time queueing system, which means that time is divided into fixed length intervals called slots. During the consecutive slots, customers arrive in the system, are stored in an infinite capacity queue and are served by a single server on a first in first out (FIFO) basis.

We assume a batch arrival process with geometrically distributed interarrival times. That is, the numbers of slots that separate consecutive slots where there are customer arrivals, constitute a series of independent and identically geometrically distributed random variables, whereas the numbers of arrivals in these consecutive arrival slots constitute a series of independent and identically distributed (i.i.d.) positive random variables with some general distribution. One easily verifies that – due to the lack of memory of the geometrical distribution – the numbers of customers arriving during the consecutive slots constitute a series of independent non-negative random variables. Therefore, one may alternatively characterise the arrival process by the common probability mass function or probability generating function of the latter series. Let thus $e(n)$ ($n \geq 0$) denote the probability that there are n arrivals during a random slot. The corresponding probability generating function $E(z)$ is then defined as

$$E(z) \triangleq \sum_{n=0}^{\infty} e(n)\, z^n. \tag{1.1}$$

One should note that the former characterisation of the arrival process does not specify arrival instants of customers within slots. Unless specified otherwise, we will not make any further assumptions.

Customer service times are synchronised on slot boundaries. That is, a customer can only start service at slot boundaries. This implies amongst other things that a customer cannot start service during his arrival slot. Further, service of a customer takes an integer number of slots, which implies that customers also leave the system at slot boundaries. The service times (in slots) of the consecutive customers are modelled by means of a series of i.i.d. positive random variables with common probability mass function $s(n)$ ($n \geq 1$) and corresponding probability generating function $S(z)$,

$$S(z) \triangleq \sum_{n=1}^{\infty} s(n)\, z^n. \tag{1.2}$$

For further use, it is convenient to introduce notation for some moments of the former distributions. In particular, using the moment generating property of probability generating functions, the mean μ_S, the variance ${\sigma_S}^2$ and the skewness γ_S of the service times relate to derivatives of the corresponding generating function as,

$$\mu_S = S'(1), \tag{1.3}$$

$${\sigma_S}^2 = S''(1) + S'(1)(1 - S'(1)), \tag{1.4}$$

$$\gamma_S = \frac{S'''(1) + [3\, S''(1) + S'(1)\,(1 - 2\, S'(1))]\,(1 - S'(1))}{(S''(1) + S'(1)(1 - S'(1)))^{\frac{3}{2}}}. \tag{1.5}$$

The mean μ_E, the variance σ_E^2 and the skewness γ_E of the arrival distribution are similarly related to the derivatives of the probability generating function $E(z)$. In general, for some random variable Z, we let μ_Z, σ_Z^2 and γ_Z denote the mean, the variance and the skewness of this variable respectively.

1.2.2 Analysis at departure epochs

Queue content

Let $V_d^{(k)}$ denote the queue content at the k-th departure epoch, that is, the number of customers in the system at the beginning of the slot following the departure slot of the k-th customer. If the k-th customer leaves a non-empty queue ($V_d^{(k)} > 0$), then a new customer starts service immediately. During this customer's service time, other customers arrive in the system and at the end of his service, this customer leaves the

system. Therefore, there are

$$V_d^{(k+1)} = V_d^{(k)} - 1 + \sum_{j=1}^{S^{(k+1)}} E^{(j)} \tag{1.6}$$

customers in the system upon departure of the $(k+1)$-th customer. Here $S^{(k+1)}$ and $E^{(j)}$ denote the service time of the $(k+1)$-th customer and the number of arrivals during the j-th slot of this service time respectively.

If the queue becomes empty upon departure of the k-th customer ($V_d^{(k)} = 0$), the server remains idle until a new customer arrives. The $(k+1)$-th customer then starts service in the slot following his arrival slot. At the end of his service time, he leaves

$$V_d^{(k+1)} = \tilde{E}^{(k+1)} - 1 + \sum_{j=1}^{S^{(k+1)}} E^{(j)} \tag{1.7}$$

customers behind in the system. Here, $\tilde{E}^{(k+1)}$ denotes the number of arrivals during the arrival slot of the $(k+1)$-th customer.

The former relations between queue content at consecutive departure epochs now translate in a relation between the corresponding probability generating functions. Let $V_d^{(k)}(z)$ denote the probability generating function of the queue content just after the k-th departure epoch, that is,

$$V_d^{(k)}(z) \triangleq \mathrm{E}\left[z^{V_d^{(k)}}\right] = \sum_{n=0}^{\infty} \Pr\left[V_d^{(k)} = n\right] z^n. \tag{1.8}$$

In the former expression, $\mathrm{E}[Z]$ and $\Pr[Z = n]$ denote the mean value of Z and the probability that Z equals n respectively. Conditioning on whether or not the queue is empty after the departure of the k-th customer then yields:

$$\begin{aligned} V_d^{(k+1)}(z) = {} & \mathrm{E}\left[z^{V_d^{(k)} - 1 + \sum_{j=1}^{S^{(k+1)}} E^{(j)}} \middle| V_d^{(k)} > 0\right] \Pr\left[V_d^{(k)} > 0\right] \\ & + \mathrm{E}\left[z^{\tilde{E}^{(k+1)} - 1 + \sum_{j=1}^{S^{(k+1)}} E^{(j)}} \middle| V_d^{(k)} = 0\right] \Pr\left[V_d^{(k)} = 0\right]. \end{aligned} \tag{1.9}$$

Due to the independence in the arrival process, the former expression further simplifies to

$$\begin{aligned} V_d^{(k+1)}(z) = {} & \frac{S(E(z))}{z} \left(V_d^{(k)}(z) - V_d^{(k)}(0)\right) \\ & + \frac{S(E(z))}{z} \tilde{E}(z)\, V_d^{(k)}(0). \end{aligned} \tag{1.10}$$

Here $\tilde{E}(z)$ denotes the probability generating function of the number of arrivals during the arrival slot of the $(k+1)$-th customer. The arrival slot of the $(k+1)$-th customer distinguishes itself from random arrival slots as there is at least one arrival during this slot. Therefore, the probability generating function $\tilde{E}(z)$ of the number of arrivals during this slot is given by

$$\tilde{E}(z) \triangleq \mathrm{E}\left[z^E | E > 0\right] = \frac{E(z) - E(0)}{1 - E(0)}. \tag{1.11}$$

The random variable E here denotes the number of customer arrivals in a random slot.

We perform a steady-state analysis. That is, we investigate the steady-state characteristics of the system under consideration. A queueing system with an infinite capacity buffer reaches steady state whenever the amount of work that arrives in the system per slot – the *load* ρ – is on average strictly less than the amount of work the system can handle per slot. That is, the system under investigation reaches steady state whenever

$$\rho = \mu_E \mu_S < 1. \tag{1.12}$$

Under this assumption, let $V_d(z) = \lim_{k\to\infty} V_d^{(k)}(z)$ denote the probability generating function of the queue content at departure epochs in steady-state. Equation (1.10) then yields,

$$V_d(z) = V_d(0)\, \frac{S(E(z))\, \left(\tilde{E}(z) - 1\right)}{z - S(E(z))}. \tag{1.13}$$

Plugging equation (1.11) into the former equation, we get

$$V_d(z) = \frac{V_d(0)}{1 - E(0)}\, \frac{S(E(z))\, (E(z) - 1)}{z - S(E(z))}. \tag{1.14}$$

The unknown factor $V_d(0)$ then follows from the normalisation property of probability generating functions, $V_d(1) = 1$:

$$V_d(0) = (1 - E(0))\, \frac{1 - \rho}{\mu_E}. \tag{1.15}$$

We used de l'Hôpital's rule to obtain the former expression. Combining the former two equations, we finally get the following expression for the probability generating function of the queue content at departure epochs in steady-state,

$$V_d(z) = \frac{1 - \rho}{\mu_E}\, \frac{E(z) - 1}{z - S(E(z))}\, S(E(z)). \tag{1.16}$$

Moments of the queue content at departure epochs in equilibrium can be retrieved using the moment generating property of probability generating functions. In particular the mean μ_{V_d} and the variance ${\sigma_{V_d}}^2$ are given by,

$$\mu_{V_d} = \frac{{\sigma_E}^2 + {\mu_E}^3 {\sigma_S}^2 - \mu_E\,(1-\rho)\,(1-\rho-\mu_E)}{2\,(1-\rho)\,\mu_E} \tag{1.17}$$

and

$${\sigma_{V_d}}^2 = \frac{\left\{\begin{array}{l} 4\,\mu_E\,(1-\rho)\left(\gamma_S {\sigma_S}^3 {\mu_E}^4 + \gamma_E {\sigma_E}^3\right) \\ +\,6\,{\mu_E}^2\left[\mu_E(2-\rho){\sigma_S}^2 + (1+\mu_S)(1-\rho)^2\right]{\sigma_E}^2 \\ +\,3\,{\mu_E}^6 {\sigma_S}^4 + 6\,{\mu_E}^4(1-\rho)^2{\sigma_S}^2 + 3\,(2\,\rho-1){\sigma_E}^4 \\ -\,{\mu_E}^2(1-\rho)^4 + {\mu_E}^4(1-\rho)^2 \end{array}\right\}}{12\,(1-\rho)^2{\mu_E}^2} \tag{1.18}$$

respectively. We again used de l'Hôpital's rule to obtain the former expressions.

A queueing system is called weakly stable, if the steady-state distribution exists, that is, if (1.12) is satisfied. This does not guarantee that this distribution possesses finite moments. If a steady-state distribution exists and if additionally this distribution also possesses a finite mean, the queueing system is called strongly stable. As can be seen from equation (1.17), the system is strongly stable if (1.12) is satisfied and if the mean and the variance of both the service times and the arrival process take finite values.

Waiting time and customer delay

Let the delay of a customer denote the number of slots between the end of the customer's arrival slot and the end of his departure slot. Similarly, let the waiting time of a customer denote the number of slots between the end of the customer's arrival slot and the beginning of the slot where this customer enters the server. One should note that both these definitions neglect the time a customer spends in the system during his arrival slot. That is, both definitions do not take the synchronisation delay into account.

As customers are served on a FIFO basis, the customers found in the queue on departure of a tagged customer are either customers that arrived during this tagged customer's arrival slot and receive service after the tagged customer or customers that arrived during the tagged customer's delay. That is, the queue content after a random customer's departure and this customer's delay are related as follows:

$$V_d = \hat{E} + \sum_{j=1}^{D} E^{(j)}. \tag{1.19}$$

Here $\hat{E}$, $E^{(j)}$ and D denote the number of arrivals during the tagged customer's arrival slot but that receive service after the tagged customer, the number of arrivals during the j-th slot of this customer's delay and the delay of this customer respectively. The former expression however does not easily translate into a corresponding expression between probability generating functions as $\hat{E}$ and D are correlated random variables.

Therefore, consider an alternative system in which all arrivals in a slot are grouped into a single batch customer with service time equal to the total service time of all these arrivals. As such, there is a batch customer arrival in a slot whenever there is at least one customer arrival in this slot. One easily observes that the number of batch arrivals during the consecutive slots and the service times of these batch customers constitute two i.i.d. series of random variables. We may therefore apply equation (1.16) to retrieve the probability generating function of the number of batch customers in the system at batch customer departure epochs in steady-state.

There is at most one batch customer arrival in a slot and this happens with probability $(1 - E(0))$. That is, there is a single batch customer arrival if there is at least one customer arrival. The common probability generating function of the numbers of batch arrivals during the consecutive slots $\mathring{E}(z)$ is then given by,

$$\mathring{E}(z) \triangleq E(0) + (1 - E(0))\, z. \tag{1.20}$$

Remember that $\tilde{E}(z)$ is the probability generating function of the number of arrivals in a slot if there is at least one arrival and therefore this is also the probability generating function of the number of customers in a batch. The batch customer service time $\mathring{S}$ equals the sum of the service times of the customers in the batch:

$$\mathring{S} = \sum_{j=1}^{\tilde{E}} S^{(j)}. \tag{1.21}$$

Here $S^{(j)}$ denotes the service time of the jth customer in a batch. The independence of the consecutive service times then implies that the probability generating function of the batch service times $\mathring{S}(z)$ is given by

$$\mathring{S}(z) = \tilde{E}(S(z)) = \frac{E(S(z)) - E(0)}{1 - E(0)}. \tag{1.22}$$

The moment generating property of probability generating functions further yields the following expressions for the mean number of batch customer arrivals in a slot $\mu_{\mathring{E}}$ and the mean batch customer service time $\mu_{\mathring{S}}$:

$$\mu_{\mathring{E}} = 1 - E(0), \tag{1.23}$$

$$\mu_{\mathring{S}} = \frac{\rho}{1 - E(0)}. \tag{1.24}$$

The probability generating function of the number of batch customers in the system upon departure of a batch customer $\mathring{V}_d(z)$, then follows from equation (1.16) and equations (1.20) to (1.24),

$$\mathring{V}_d(z) = \frac{(1-\rho)\,(z-1)\,(E(S(E(0)+(1-E(0))\,z))-E(0))}{E(0)+(1-E(0))\,z-E(S(E(0)+(1-E(0))\,z))}. \tag{1.25}$$

There is at most one batch customer arrival per slot. Therefore, the number of batch customers in the queue upon departure of a random batch customer equals the number of batch customers that arrived during the departing batch customer's delay. The number of batch customers $\mathring{V}_d$ in the queue upon departure of a random batch customer and this batch customer's delay $\mathring{D}$ are therefore now related as

$$\mathring{V}_d = \sum_{j=1}^{\mathring{D}} \mathring{E}^{(j)}. \tag{1.26}$$

Here $\mathring{E}^{(j)}$ denotes the number of batch customer arrivals during the j-th slot of the tagged batch customer's delay. The former now easily translates into,

$$\mathring{V}_d(z) = \mathring{D}(\mathring{E}(z)) = \mathring{D}(E(0)+(1-E(0))\,z), \tag{1.27}$$

or equivalently, after plugging in equation (1.25),

$$\mathring{D}(z) = \frac{1-\rho}{1-E(0)}\,(z-1)\,\frac{E(S(z))-E(0)}{z-E(S(z))}. \tag{1.28}$$

Here $\mathring{D}(z)$ denotes the probability generating function of the batch customer delay.

By definition, the batch customer delay is the sum of the batch customer waiting time $\mathring{W}$ and the batch customer service time $\mathring{S}$,

$$\mathring{D} = \mathring{W} + \mathring{S}. \tag{1.29}$$

Since $\mathring{W}$ and $\mathring{S}$ are independent random variables, the probability generating function of the waiting time of a batch customer $\mathring{W}(z)$ is given by,

$$\mathring{W}(z) = \frac{\mathring{D}(z)}{\mathring{S}(z)} = (1-\rho)\,\frac{z-1}{z-E(S(z))}. \tag{1.30}$$

We are now ready to reconsider the customer delay in the original system. A random (tagged) customer's delay D is the sum of the waiting time $\mathring{W}$ of the corresponding batch customer, of the service times $S^{(j)}$ of all customer arrivals during the tagged

customer's arrival slot that are served before this tagged customer and of this tagged customer's service time S:

$$D = \mathring{W} + \sum_{j=1}^{\check{E}} S^{(j)} + S. \tag{1.31}$$

Here $\check{E}$ denotes the number of customer arrivals during the tagged customer's arrival slot that receive service before this tagged customer. The independence in the arrival and service processes then yields the following expression for the probability generating function of the delay of a random customer $D(z)$,

$$D(z) = \mathring{W}(z)\, \check{E}(S(z))\, S(z), \tag{1.32}$$

where $\check{E}(z)$ denotes the probability generating function corresponding to $\check{E}$. The latter can be retrieved as follows.

First consider the number of arrivals during a random customer's arrival slot. Based on a simple counting argument, one retrieves following expression for the probability $\dot{e}(n)$ ($n \geq 1$) that there are n arrivals during a random customer's arrival slot,

$$\dot{e}(n) = \frac{n\, e(n)}{\mu_E}. \tag{1.33}$$

Clearly, the probability mass functions of the numbers of arrivals during a random slot and during the arrival slot of a random customer are not equal. This is a manifestation of the so-called renewal-theory paradox [Cooper et al., 1997]. The probability $\check{e}(n)$ ($n \geq 0$) that there are n customer arrivals in a random customer's arrival slot that are served before this random customer is then retrieved by conditioning on the total number of arrivals during this random customer's arrival slot:

$$\check{e}(n) = \sum_{i=n+1}^{\infty} \frac{1}{i}\, \dot{e}(i). \tag{1.34}$$

After plugging equation (1.33) into the former expression, one easily retrieves the corresponding probability generating function:

$$\check{E}(z) = \frac{E(z) - 1}{\mu_E\, (z - 1)}. \tag{1.35}$$

Plugging equations (1.30) and (1.35) into equation (1.32), then finally yields

$$D(z) = \frac{1 - \rho}{\mu_E}\, \frac{E(S(z)) - 1}{z - E(S(z))}\, \frac{(z - 1)\, S(z)}{S(z) - 1}. \tag{1.36}$$

By means of the moment generating property of probability generating functions, we can retrieve the various moments of the customer delay. In particular, the mean customer delay μ_D and the corresponding variance σ_D^2 are given by,

$$\mu_D = \frac{\rho(1-\rho) + {\sigma_S}^2{\mu_E}^2 + \mu_S{\sigma_E}^2}{2\,(1-\rho)\,\mu_E} \tag{1.37}$$

and

$${\sigma_D}^2 = \frac{\left\{\begin{array}{l} -\,(1-\rho)^4 + 2\,(1-\rho)^3(1-{\mu_E}^2) \\ -\,(1-\rho)^2(4\,\gamma_E{\sigma_E}^3\mu_S - 2\,{\mu_E}^2 + 1 - 6\,{\sigma_S}^2{\mu_E}^2) \\ +\,2\,(1-\rho)(2\,\gamma_S{\sigma_S}^3{\mu_E}^3 + 2\,\gamma_E{\sigma_E}^3\mu_S - 3\,{\sigma_E}^4{\mu_S}^2) \\ +\,6\,{\sigma_E}^2{\sigma_S}^2\mu_E + 3\,{\sigma_E}^4{\mu_S}^2 + 3\,{\sigma_S}^4{\mu_E}^4 \end{array}\right\}}{12\,{\mu_E}^2\,(1-\rho)^2} \tag{1.38}$$

respectively.

Similarly as for the batch customer waiting time, we relate the probability generating function of the customer waiting time $W(z)$ to the probability generating function of the customer delay:

$$W(z) = \frac{D(z)}{S(z)}. \tag{1.39}$$

Mean μ_W and variance σ_W^2 of the customer waiting time then relate to mean and variance of customer delay as follows:

$$\mu_W = \mu_D - \mu_S, \tag{1.40}$$

$${\sigma_W}^2 = {\sigma_D}^2 - {\sigma_S}^2. \tag{1.41}$$

1.2.3 Analysis at random slot boundaries

Joint probability generating function

At a random slot boundary in steady-state, the *state* of the system is characterised by the number of customers in the queue V_r and by the remaining service time H of the customer that receives service. We here use the term "state" in the Markovian sense. That is, given the state, future queueing behaviour does not depend on past behaviour. If a customer receives service in the slot following the observed slot boundary, the remaining service time is defined as the number of slots between this slot boundary and the end of the slot where the customer in service leaves the system. Further, the remaining service time equals 0 slots by definition if there are no customers in the

system at the observed slot boundary. Notice that H only takes positive values when there are customers present at the observed slot boundary. We now concentrate on the joint probability generating function $P(x, z)$ of these variables in steady-state:

$$P(x, z) \triangleq \mathrm{E}\left[x^H \, z^{V_r}\right]. \tag{1.42}$$

One easily verifies that the server is occupied for a fraction ρ of the slots. Therefore, conditioning on whether or not there is a customer in service yields

$$P(x, z) = (1 - \rho) + \rho \, \hat{P}(x, z), \tag{1.43}$$

where $\hat{P}(x, z)$ denotes the joint probability generating function of the queue content and the remaining service time at the beginning of a slot where a customer receives service (a *busy slot*):

$$\hat{P}(x, z) \triangleq \mathrm{E}\left[x^H \, z^{V_r} | H > 0\right]. \tag{1.44}$$

Let V_s denote the queue content at a random start epoch in steady-state. That is, V_s denotes the queue content at the beginning of a slot where a customer starts service. Assume that this customer needs S slots of service. At the beginning of the i-th service slot of this customer, the remaining service time then equals $S - i + 1$ slots. Further, the queue content at the beginning of this slot equals the queue content at the beginning of the slot where this customer starts service, augmented by the customer arrivals during the first $i - 1$ service slots. Using renewal-reward theory (see [Takagi, 1991, pp. 202–205]), these observations yield

$$\hat{P}(x, z) = \frac{1}{\mu_S} \mathrm{E}\left[\sum_{i=1}^{S} x^{S-i+1} \, z^{V_s + \sum_{j=1}^{i-1} E^{(j)}}\right]. \tag{1.45}$$

Here $E^{(j)}$ denotes the number of arrivals in the system during the j-th service slot of this customer. One may also obtain the former expression by conditioning on the total service time and on the remaining service time of the customer that receives service during a random busy slot. Expression (1.45) now further simplifies to,

$$\hat{P}(x, z) = \frac{x \, V_s(z)}{\mu_S} \, \frac{S(E(z)) - S(x)}{E(z) - x}, \tag{1.46}$$

where $V_s(z)$ denotes the probability generating function of the queue content at start epochs.

When a customer departs from the system, he leaves behind all customers that were present in the system when this customer started service, augmented with all arrivals

during his service time and diminished by 1 as the customer himself leaves the system. Therefore, the queue content at start and departure epochs are related as,

$$V_d = V_s - 1 + \sum_{j=1}^{S} E^{(j)}. \tag{1.47}$$

The latter then translates into the following relation between the corresponding probability generating functions,

$$V_s(z) = \frac{z}{S(E(z))}\, V_d(z). \tag{1.48}$$

In view of the former expression and equation (1.16), equation (1.46) simplifies to

$$\hat{P}(x,z) = \frac{1-\rho}{\rho} xz \frac{S(E(z)) - S(x)}{z - S(E(z))}\, \frac{E(z)-1}{E(z)-x}. \tag{1.49}$$

Plugging the latter expression into equation (1.43) then yields the following expression for the joint probability generating function $P(x,z)$,

$$P(x,z) = (1-\rho)\left(1 - xz\,\frac{S(E(z)) - S(x)}{S(E(z)) - z}\,\frac{1-E(z)}{x-E(z)}\right). \tag{1.50}$$

Bruneel and Kim [1993] retrieve the former equation directly from a set of system equations. Other generating functions are then derived from this joint probability generating function.

Queue content, unfinished work and customer delay

We may retrieve the probability generating function of the queue content at random slot boundaries in steady-state $V_r(z)$ by evaluation of $P(x,z)$ for $x=1$,

$$V_r(z) = P(1,z) = (1-\rho)\,\frac{(z-1)\,S(E(z))}{z - S(E(z))}. \tag{1.51}$$

The moment generating property of the generating function then yields expressions for the mean μ_{V_r} and the variance ${\sigma_{V_r}}^2$ of the queue content at random slot boundaries. Mean and variance are given by

$$\mu_{V_r} = \frac{\rho(1-\rho) + {\sigma_S}^2\,{\mu_E}^2 + \mu_S\,{\sigma_E}^2}{2\,(1-\rho)} \tag{1.52}$$

and

$$\sigma_{V_r}{}^2 = \frac{\left\{\begin{array}{l} -(1-\rho)^4 + \left(6\,\sigma_E{}^2\mu_S + 1 + 6\,\sigma_S{}^2\mu_E{}^2\right)(1-\rho)^2 \\ + \left(4\,\gamma_E\sigma_E{}^3\mu_S + 6\,\sigma_S{}^2\sigma_E{}^2\mu_E + 4\,\gamma_S\sigma_S{}^3\mu_E{}^3\right)(1-\rho) \\ + \left(3\,\sigma_S{}^4\mu_E{}^4 + 6\,\sigma_S{}^2\sigma_E{}^2\mu_E + 3\,\sigma_E{}^4\mu_S{}^2\right) \end{array}\right\}}{12\,(1-\rho)^2} \tag{1.53}$$

respectively.

Let the *unfinished work* denote the number of slots it takes to return to an empty system under the assumption that there are no new customer arrivals. Clearly, if there is at least one customer present in the system ($H > 0$), the unfinished work at the beginning of a random slot U consists of the service times $S^{(j)}$ ($j = 1 \ldots V_r - 1$) of all customers waiting in the queue, augmented with the remaining service time H of the customer in service. That is,

$$U = H + \sum_{j=1}^{V_r-1} S^{(j)}. \tag{1.54}$$

Further, if the system is empty, the unfinished work U equals by definition 0 slots. The probability generating function of the unfinished work at random slot boundaries $U(z)$ then easily follows from (1.50),

$$\begin{aligned} U(z) &= \Pr[H=0] + \mathrm{E}\left[z^{H+\sum_{j=1}^{V_r-1} S^{(j)}} | H>0\right] \Pr[H>0] \\ &= P(0,0) + (P(z,S(z)) - P(0,0))\,\frac{1}{S(z)} \\ &= (1-\rho)\frac{E(S(z))(z-1)}{z-E(S(z))}. \end{aligned} \tag{1.55}$$

The moment generating property of probability generating functions then yields the following expressions for the mean μ_U and the variance $\sigma_U{}^2$ of the unfinished work at random slot boundaries:

$$\mu_U = \frac{\mu_S{}^2\sigma_E{}^2 + \mu_E\sigma_S{}^2 + \rho(1-\rho)}{2\,(1-\rho)}, \tag{1.56}$$

$$\sigma_U{}^2 = \frac{\left\{\begin{array}{l} 4\,\left(\mu_E\gamma_S\sigma_S{}^3 + \mu_S{}^3\gamma_E\sigma_E{}^3\right)(1-\rho) + 6\,\mu_S{}^2\sigma_E{}^2(1-\rho)^2 \\ + 6\,\left[\left(\sigma_E{}^2\mu_S - \rho\mu_E\right)(2-\rho) + \mu_E\right]\sigma_S{}^2 \\ + \rho(2-\rho)(1-\rho)^2 + 3\,\left(\mu_E{}^2\sigma_S{}^4 + \mu_S{}^4\sigma_E{}^4\right) \end{array}\right\}}{12\,(1-\rho)^2}. \tag{1.57}$$

We may now again retrieve the probability generating function of the customer delay, this time using knowledge of the unfinished work at random slot boundaries. Consider a random (tagged) customer. His delay consists of the unfinished work at the beginning of his arrival slot diminished by one if there is such unfinished work, the service times of all customers that arrived during this tagged customer's arrival slot and that are served before the tagged customer and the tagged customer's service time,

$$D = (U-1)^+ + \sum_{j=1}^{\check{E}+1} S^{(j)}. \tag{1.58}$$

Here $\check{E}$ and $S^{(j)}$ denote the number of customer arrivals in the tagged customer's arrival slot but that are served before this customer and the service time of the j-th customer in the tagged customer's arrival slot. Further, the notation $(.)^+$ is shorthand for $\max(., 0)$. The independence in the arrival process implies that the system state at a random customer's arrival slot is statistically indistinguishable from the system state at a random slot boundary. Therefore, (1.55) also denotes the probability generating function of the unfinished work at the beginning of a random customer's arrival slot. The former equation then yields the following expression for the probability generating function of the customer delay:

$$D(z) = \left((U(z) - U(0))\frac{1}{z} + U(0) \right) \check{E}(S(z))S(z). \tag{1.59}$$

After substitution of equations (1.35) and (1.55), we again retrieve equation (1.36).

1.2.4 Analysis at arrival epochs

Observation of queue sizes at arrival epochs requires additional assumptions regarding the exact arrival times of customers within slots. Upon arrival, a (tagged) customer finds the customers that were present in the system at the beginning of the customer's arrival slot, augmented with all customers that arrived in the same slot but at a time before the tagged customer's arrival time. Notice that the tagged customer does not find customers that arrive at the same time instant as the tagged customer upon arrival. Therefore, the number of customers that a tagged customer finds upon arrival does not necessarily include all customers that are served before the tagged customer.

We here limit our discussion to the case that all customers arrive at distinct epochs within slots. Under this assumption, the tagged customer finds all customers that are served before this customer – recall that we assume a FIFO discipline – upon arrival. Therefore, the queue content at arrival epochs V_a and queue content at random slot boundaries V_r are related as

$$V_a = V_r + \check{E}. \tag{1.60}$$

As before, $\check{E}$ denotes the number of customers during a tagged customer's arrival slot that are served before this customer. The corresponding probability generating function is displayed in equation (1.35). The independence in the arrival process implies that the queue content at the beginning of a random customer's arrival slot and the queue content at the beginning of a random slot are statistically indistinguishable. The former expression then yields the following expression for the probability generating function of the queue content at arrival epochs:

$$V_a(z) = V_r(z)\,\check{E}(z). \tag{1.61}$$

It is now easy to verify that the generating functions of the queue content at arrival and departure epochs are equal:

$$V_a(z) = V_d(z). \tag{1.62}$$

This result also directly follows from Burke's theorem (see [Takagi, 1991, pg. 7]). The latter theorem applies to queueing systems with neither batch arrivals nor batch departures. That is, equation (1.62) is valid for any single server queueing system with arrivals that occur at distinct instants within slots.

Combining Burke's theorem and equation (1.61), one further observes that the probability generating functions of the queue content at departure epochs and at random slot boundaries are related as

$$V_d(z) = V_r(z)\,\check{E}(z). \tag{1.63}$$

In accordance with Burke's theorem and with the assumptions that led to equation (1.61), this expression is valid for any single server queueing system with independent arrivals that occur at distinct instants within slots. Further, neither queue content at departure epochs nor at random slot boundaries depend on the exact timing of arrivals within slots. Therefore (1.63) remains valid in the case of arrival bursts during slots. That is, equation (1.63) is valid for any discrete-time single server queueing system with an independent arrival process.

1.2.5 Little's result

Comparison of equations (1.37) and (1.52) yields the surprisingly simple relation,

$$\mu_{V_r} = \mu_E\,\mu_D, \tag{1.64}$$

between the mean (discrete-time) customer delay and the mean queue content at random slot boundaries. This result is remarkably similar to Little's result (see e.g., Little [1961], Whitt [1991]) which relates the mean (continuous-time) customer delay $\mu_{\mathcal{D}}$

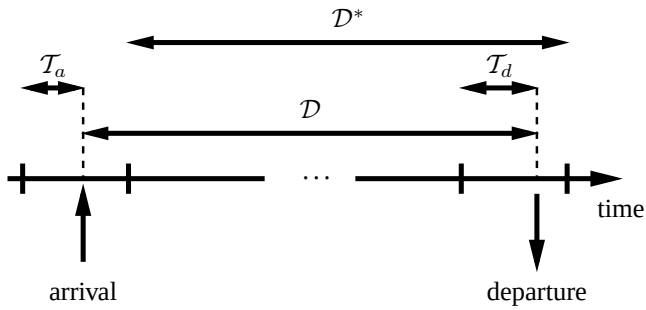


Figure 1.3: A customer's delay in the original system and in the system with rescheduling.

and the mean queue content at random points in time $\mu_{\mathcal{V}_r}$ for general continuous-time queueing systems,

$$\mu_{\mathcal{V}_r} = \lambda\, \mu_{\mathcal{D}}. \tag{1.65}$$

Here λ denotes the arrival intensity, that is,

$$\lambda = \lim_{t\to\infty} \frac{\mathrm{E}[\mathcal{E}(t)]}{t}, \tag{1.66}$$

where $\mathcal{E}(t)$ equals the number of customer arrivals in the interval $[0, t]$. One often refers to equation (1.64) as Little's result as well. However, different quantities are related. In particular, this discrete-time equivalent result depends on the definition of the discrete-time delay introduced in section 1.2.2, as we will see further. We first focus on a more general discrete-time equivalent of Little's result. The argument closely follows our contribution [Fiems and Bruneel, 2002b].

Discretisation of Little's result

Consider a general discrete-time queueing system and let μ_{V_r} denote the mean queue content observed at random slot boundaries. The exact arrival and departure times – as long as one retains the order in which these events occur – within slots do not influence the queue content observed at slot boundaries. Therefore, one observes the same mean queue content at random slot boundaries for an alternative system where all arrivals and departures are rescheduled to (just before) slot boundaries. In this new system, the queue content only changes at slot boundaries, and therefore μ_{V_r} also denotes the mean queue content at random points in time for the system with rescheduling.

Comparison of the delay in the original system and the system where arrivals and

departures are rescheduled then yields

$$\mu_{\mathcal{D}} = \mu_{\mathcal{D}^*} - \mu_{\mathcal{T}_a} + \mu_{\mathcal{T}_d}. \tag{1.67}$$

Here $\mu_{\mathcal{D}}$ denotes the mean customer delay in the system under consideration, $\mu_{\mathcal{D}^*}$ denotes the mean customer delay in the system with rescheduling and $\mu_{\mathcal{T}_a}$ and $\mu_{\mathcal{T}_d}$ denote the mean arrival time and the mean departure time of a random customer relative to the preceding slot boundary. Figure 1.3 illustrates the relation between a customer's arrival time $\mathcal{T}_a$ and departure time $\mathcal{T}_d$ relative to the preceding slot boundary and the customer delays $\mathcal{D}$ and $\mathcal{D}^*$ in the original system and the system with rescheduling. The former expression then easily follows. Note that the term customer delay here refers to the continuous-time customer delay, i.e., the time between a customer's arrival epoch and his departure epoch.

Clearly, Little's (continuous-time) result applies to the system with rescheduling and therefore

$$\mu_{V_r} = \lambda\,(\mu_{\mathcal{D}} + \mu_{\mathcal{T}_a} - \mu_{\mathcal{T}_d}). \tag{1.68}$$

Although equation (1.68) does relate the mean continuous-time customer delay and the mean queue content at random slot boundaries, direct use in discrete time is rather limited due to the presence of $\mu_{\mathcal{T}_a}$ and $\mu_{\mathcal{T}_d}$.

Definition of discrete-time delay

It now turns out that it is the definition of the "discrete-time delay" that leads to the applicable relation (1.64). Remember that the discrete-time delay is defined as the number of slots between the end of a customer's arrival slot and the end of his departure slot. Therefore, one easily establishes that the mean continuous-time delay $\mu_{\mathcal{D}}$ and the mean discrete-time delay μ_D are related as follows:

$$\mu_{\mathcal{D}} = \mu_D \Delta - \mu_{\mathcal{T}_a} + \mu_{\mathcal{T}_d}. \tag{1.69}$$

Here Δ denotes the slot length. Further, given the slot length, the mean number of arrivals per slot and the arrival intensity are related as,

$$\mu_E = \lambda\,\Delta. \tag{1.70}$$

Substitution of the former expressions in equation (1.68) then yields (1.64). That is, the discrete-time equivalent of Little's result (1.64) is applicable as long as one adheres to the definition of the discrete-time delay.

1.3 Overview

To conclude this chapter, we will give an outline of the following chapters that concentrate on discrete-time queueing systems with vacations.

In chapter 2, queueing models with random vacations are investigated. The notion "random" here refers to the fact that vacations occur independently of the state (queue content, remaining customer service time, ...) of the system. As this in particular implies that vacations may interrupt the service of a customer, one may also refer to these systems as systems with service interruptions. Different models are proposed with varying complexity regarding the considered vacation processes as well as regarding the operation modes to cope with interrupted service. The vacations are modelled as a Bernoulli process, as a Markovian on-off process and as an on-off process with geometrically distributed on-times and generally distributed off-times. For all these vacation models we look into the following three operation modes to cope with service interruptions: the continue after interruption mode, the repeat after interruption mode and the repeat after interruption with resampling mode. For the Bernoulli vacation model, we also investigate some variants of the former operation modes: delayed modes and partial modes. As an application, we investigate a preemptive multi-class priority queueing system.

As opposed to chapter 2, the vacation processes of the models of chapter 3 take the state of the system into account. Classical vacation models of this type include amongst others the exhaustive and the gated vacation models. The proposed gated-exhaustive vacation model encapsulates both these vacation systems. We further study a fairly general vacation model that encapsulates most classical non-gated vacation queueing systems.

Finally, we conclude this dissertation by summarising the main results in chapter 4.

Chapter 2

Random server vacations

The queueing systems that are considered in the present chapter, share the property that server vacations occur independently of the state of the queueing system. That is, vacations may occur whether or not there are customers in the queue, whether or not a customer is being served. This implies in particular that the server can leave for a vacation while there is a customer in service. That is, the service of a customer can be interrupted. Therefore, these models are also referred to as (server) interruption models or queueing systems with (service) interruptions. Before the presentation of our results, we briefly survey related literature. We consider literature on both discrete-time and continuous-time queueing systems with interruptions.

We first consider continuous-time contributions. According to Ibe and Trivedi [1990], White and Christie [1958] were the first to study queues with interruptions. They consider an $M/M/1$ queueing system where vacations are modelled as an on-off process with exponentially distributed on- and off-periods. Generally distributed service times and off-periods are considered by Avi-Itzhak and Naor [1963] and also by Thiruvengadam [1963]. These authors consider exponentially distributed on-periods as opposed to Federgruen and Green [1986], who consider phase-type on-periods. Van Dijk [1988] provides an approximate analysis of a system with exponentially distributed service times but with generally distributed on- and off-periods whereas Takine and Sengupta [1997] study a vacation queueing system in a Markov-modulated environment. The latter authors also allow correlation in the arrival process. Further, a processor sharing queueing system with exponentially distributed on-periods and generally distributed off-periods is studied by Núñez Queija [2000]. All these contributions assume that customers resume service after the interruption. Gaver Jr. [1962] also considers the cases where service is repeated or repeated and resampled after the interruption. The latter operation mode is also studied by Ibe and Trivedi [1990] for a two station polling system.

Research on discrete-time queueing systems with random server vacations started

later. Early contributions include those by Hsu [1974] and Heines [1979]. Both authors treat the single server system with Bernoulli server vacations and a Poisson arrival process. The former considers queue content at random slot boundaries whereas the latter considers queue content at service completion times. A single server system with an independent arrival process and a correlated on/off server vacation process is treated by Bruneel [1986], by Yang and Mark [1990] and by Woodside and Ho [1987]. Woodside and Ho [1987] and Yang and Mark [1990] model the on and off-periods as two series of independent shifted geometric random variables, whereas Bruneel [1986] assumes that the series of consecutive on-periods as well as the series of consecutive off-periods share a common general distribution. The only restriction in the latter contribution is that the common probability generating function of the on-periods must be rational. Alternatively, correlation in the vacation process is captured by means of a Markovian process by Lee [1997a]. In a more general setting – that is, no assumptions are made regarding the nature of the vacation process – relationships between queue content at different time epochs are derived by Bruneel [1983b].

Georganas [1976] and Bruneel [1984a] treat multi-server systems with independent customer arrival and server vacation processes. The latter extends the former in the sense that it does not assume that all servers are either available or on vacation simultaneously. The delay analysis of the latter system is presented by Laevens and Bruneel [1995]. Bruneel [1985] also considers a multi-server system with a correlated vacation process. Here, the vacation process is modelled as an on/off process (geometrical on-periods). The numbers of available servers during the consecutive on-slots constitute a series of i.i.d. non-negative random variables whereas no servers are available during off-periods.

Some contributions also allow a certain degree of correlation in the arrival process. Bruneel [1984b] assumes that both arrival and vacation processes are on/off processes with geometric on- and off-periods. A stochastic number of customers enters the system during arrival-on periods, whereas no customers arrive in the system during arrival-off periods. The vacation process is similar as the one analysed by Yang and Mark [1990] in the case of uncorrelated arrivals. This vacation process is also considered by Ali et al. [2001]. These authors however assume that customer arrivals come from a superposition of two-state Markovian on-off sources.

All the former discrete-time queueing models have fixed customer service times of a single slot in common. A queueing system where customers have fixed multiple-slot service times, is considered by Inghelbrecht et al. [2000]. The vacation process is again similar to the one treated by Yang and Mark [1990]. The presence of multiple-slot service times and random vacations implies – similarly as for continuous-time models – that the server may take a vacation while a customer receives service. The paper considers both the case that a customer's service is resumed after the interruption and the case that service is repeated after the interruption. Systems with more general service time distributions and different interruption models are the subject of this chapter.

The outline of the rest of this chapter is as follows. The following two sections consider queueing systems subjected to Bernoulli vacations. Different analysis methods are presented and compared. In section 2.3, we investigate a queueing system subjected to a two-state Markovian vacation process. Our results are extended in section 2.4 to queueing systems with geometrically distributed available periods and generally distributed vacations. The theoretical results are then applied to analyse a discrete-time multi-class preemptive priority system in section 2.5. We focus on an exact analysis as well as on approximations. Sections 2.1 to 2.4 all focus on three different operation modes to handle interrupted service: the continue after interruption mode, the repeat after interruption mode and the repeat after interruption mode with resampling. In section 2.6, we investigate some other operation modes. We here again assume a Bernoulli vacation process.

2.1 Bernoulli vacations

In this section, we consider the discrete-time $Geo^X/G/1$ queue subjected to Bernoulli vacations. Results are based on our contribution [Fiems et al., 2001], and also on our contribution [Fiems and Bruneel, 2002b]. However, the analysis, as presented in the following paragraphs, is somewhat different.

2.1.1 Queueing model

As before, the numbers of customer arrivals during the consecutive slots are modelled by means of a series of i.i.d. non-negative random variables. The arrival stream is then characterised by the common probability mass function $e(n)$ ($n \geq 0$) or the corresponding probability generating function $E(z)$ of this series. Further, service times of the consecutive customers constitute a series of i.i.d. positive random variables with common probability mass function $s(n)$ ($n > 0$) and corresponding probability generating function $S(z)$. Service of customers is synchronised with respect to slot boundaries. Recall that this implies that a customer cannot commence service during his arrival slot.

There is a single server which is not always available. The numbers of available (0 or 1) servers during the consecutive slots are modelled by means of an independent and identically Bernoulli distributed series of random variables, characterised by the probability σ ($0 < \sigma \leq 1$) that the server is available during a random slot.

Slots during which the server is available are called A-slots or available slots, and analogously, slots during which the server takes a vacation are called B-slots or blocked slots. An A-period (B-period) is defined as a number of contiguous slots during which the server is continuously available (on vacation). Due to the Bernoulli nature of the

vacation process, the series of consecutive A- and B-periods are both series of independent and identically geometrically distributed random variables with parameters σ and $1-\sigma$ respectively. Their common probability generating functions are given by

$$A(z) = \frac{(1-\sigma)z}{1-\sigma z} \tag{2.1}$$

and

$$B(z) = \frac{\sigma z}{1-(1-\sigma)z} \tag{2.2}$$

respectively.

As customer service times take in general more than one slot, a B-period may start during a customer's service time. We investigate different ways, say operation modes, to handle interrupted service. In the *continue after interruption* (CAI) mode, a customer's service continues after a server interruption, i.e., service resumes with the part of the interrupted service time that has not been completed yet. In the *repeat after interruption* (RAI) mode, the complete customer service time is repeated if an interruption occurs. This is also the case for the *repeat after interruption with resampling* (RAI,wr) mode. However for this operation mode, service times are resampled after each interruption.

One may note that RAI and RAI,wr modes are equivalent if customer service times have a fixed length. Also, CAI and RAI,wr operate equivalently in the particular case of geometrically distributed service times. The latter is due to the absence of memory of the geometric distribution. Further, all modes operate equivalently when the customer service time is deterministically equal to one slot or when there are no server vacations. In either case, service of a customer is never interrupted.

2.1.2 Effective service times

The effective service time of a customer is defined as the number of slots required to serve a customer. It is the time period starting with the slot where the customer enters the server and ending with the slot where the customer leaves the system. A customer enters the server at the beginning of the slot following his arrival slot if he is the first customer of a batch entering an empty system or at the beginning of the slot following the departure slot of the preceding customer if this is not the case. Note that entering the server at the beginning of a slot does not mean that the customer receives service during this slot as the server may be on vacation. Figures 2.1 to 2.3 illustrate this definition. The effective service times in case of CAI, RAI and RAI,wr mode are depicted for a customer with an original customer service time of 5 slots. In case of RAI,wr the latter is resampled to 4 slots after the interruption. Note in particular,

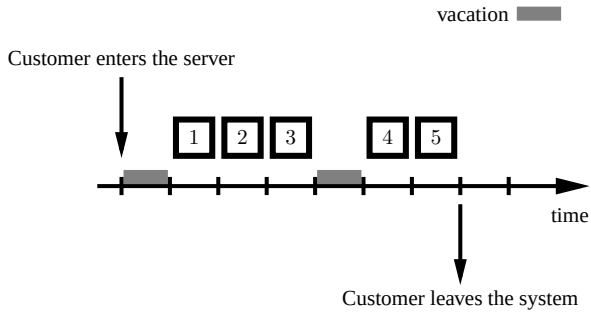


Figure 2.1: Effective service time for the continue after interruption mode.

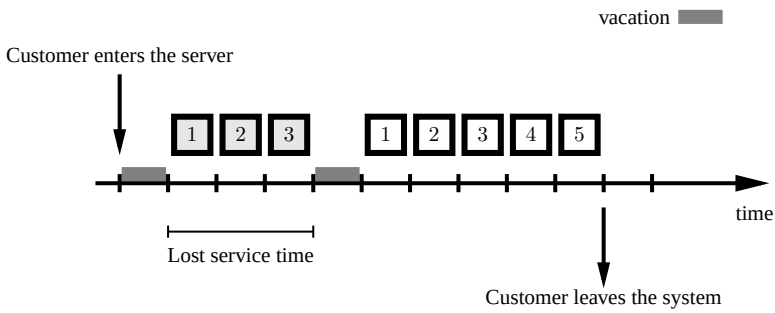


Figure 2.2: Effective service time for the repeat after interruption mode.

that the effective service times also include the B-slots preceding the slot where the customer receives service for the first time.

The Bernoulli nature of the vacation process and the i.i.d. nature of the consecutive service times imply that the effective service times of the consecutive customers constitute a series of i.i.d. random variables. We denote the corresponding common probability generating function of this series by $T(z)$. Expressions for the latter probability generating function are retrieved in the following subsections for all operation modes under consideration.

Continue after interruption

In case of the CAI operation mode, the customer receives service immediately if the server is available during the first slot of his effective service time, that is, if this first slot is an A-slot. Further, the customer receives for the first time service during the n-th effective service time slot if the server is blocked during the first $(n-1)$ slots and available during the n-th slot. The Bernoulli nature of the interruption process then

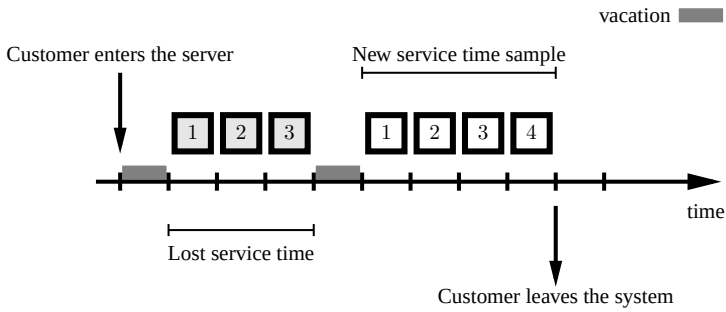


Figure 2.3: Effective service time for the repeat after interruption with resampling mode.

yields that the probability that the customer is first served during the n-th slot of his effective service time is given by,

$$\psi(n) = (1-\sigma)^{n-1}\sigma. \tag{2.3}$$

Similarly, the customer has to wait another $(n-1)$ slots after he has received service before he receives service again with probability $\psi(n)$. Therefore the effective service time of customer with service time S is the sum of S independent random variables with common probability mass function given by (2.3) or with corresponding probability generating function given by,

$$\Psi(z) = \frac{\sigma z}{1-(1-\sigma)z}. \tag{2.4}$$

The probability generating function of the effective service time in case of CAI, is then the probability generating function of a stochastic sum of i.i.d. random variables, i.e.,

$$T(z) = S\left(\Psi\left(z\right)\right) = S\left(\frac{\sigma z}{1-(1-\sigma)z}\right). \tag{2.5}$$

The moment generating property of probability generating functions then allows us to obtain various moments of the effective service times in case of CAI mode. In particular, we obtain the following expressions for the mean μ_T, the variance ${\sigma_T}^2$ and the skewness γ_T,

$$\mu_T = \frac{\mu_S}{\sigma}, \tag{2.6}$$

$${\sigma_T}^2 = \frac{{\sigma_S}^2 + (1-\sigma)\mu_S}{\sigma^2}, \tag{2.7}$$

and,

$$\gamma_T = \frac{\gamma_S {\sigma_S}^3 + 3\,(1-\sigma){\sigma_S}^2 + (1-\sigma)(2-\sigma)\mu_S}{\sigma^3 {\sigma_T}^3}. \tag{2.8}$$

Recall that expressions for mean, variance and skewness in terms of derivatives of the corresponding probability generating function can be found in chapter 1 (equations (1.3), (1.4) and (1.5) respectively).

In general, the n-th moment of the effective service time will be a function of the vacation parameter σ and the first n moments of the customer service time distribution. Moments of the effective service times therefore exist whenever the corresponding moments of the service times exist for all $0 < \sigma \leq 1$.

Alternatives to the current analysis methodology are presented in [Fiems et al., 2001] and in [Fiems et al., 2002c]. The former contribution notes that the effective service time equals n slots given that the service time equals $k < n$ slots, if the server is available during the last slot of the effective service time and during $(k-1)$ slots of the preceding effective service time slots. The latter contribution derives a recursive relation by conditioning on the availability of the server during the first slot of the effective service time. This methodology is also used in some of the subsequent sections.

Repeat after interruption

In case of RAI operation, we first consider the effective service times given some fixed service time, say $S = k$. The effective service time equals the service time of the customer if no interruptions occur or equals the sum of the length of an unsuccessful attempt and the remaining effective service time after this attempt if this is not the case:

$$T = \begin{cases} S & \text{no interruptions,} \\ \Phi + \tilde{T} & \text{unsuccessful attempt.} \end{cases} \tag{2.9}$$

Here T and $\tilde{T}$ denote the effective service time and the remaining effective service time after an unsuccessful attempt of a customer respectively. Φ denotes the length of an unsuccessful service attempt and S denotes the service time of the customer. An unsuccessful attempt starts at the beginning of the slot where the customer enters the server and ends at the end of the next B-slot under the assumption that the customer did not finish his service by then. Notice that an unsuccessful attempt takes a single slot if the customer enters the server at the beginning of a B-slot.

Given the customer service time, one observes that Φ and $\tilde{T}$ are independent random variables. This is due to the nature of the interruption process. Further – given the service time – the remaining effective service time $\tilde{T}$ and the effective service time T

share a common probability generating function as service has to start all over after the attempt. Let $T(z|k)$ denote the probability generating function of the (remaining) effective service time of a customer that requires k slots of service. An attempt is successful if the server remains available for k or more consecutive slots. The i.i.d. nature of the vacation process implies that the server remains available during k consecutive slots with probability σ^k. Therefore, we get,

$$T(z|k) = \sigma^k z^k + (1 - \sigma^k)\Phi(z|k)T(z|k) \tag{2.10}$$

where $\Phi(z|k)$ denotes the conditional probability generating function of the number of slots of an unsuccessful attempt given that the customer service time equals k slots.

Given the customer service time $S = k$, an attempt takes j ($0 < j \leq k$) slots, if the server remains available for $j - 1$ slots and becomes unavailable. Let $\phi(j|k)$ denote the probability that the attempt takes j slots, given that the service time takes k slots, then,

$$\phi(j|k) = (1 - \sigma)\frac{\sigma^{j-1}}{1 - \sigma^k} \tag{2.11}$$

for $0 < j \leq k$ and $\phi(j|k)$ equals 0 elsewhere. The corresponding conditional probability generating function is then easily derived:

$$\Phi(z|k) = \frac{(1 - \sigma)z}{1 - \sigma z}\frac{1 - (\sigma z)^k}{1 - \sigma^k}. \tag{2.12}$$

Substitution of the former expression in equation (2.10) then yields following expression for the probability generating function of the effective service times given the customer service time equals k slots:

$$T(z|k) = \frac{(\sigma z)^k(1 - \sigma z)}{1 - z + (1 - \sigma)z(\sigma z)^k}. \tag{2.13}$$

Summation over all possible service times $S = k$ then finally yields an expression for the probability generating function of the effective service times,

$$T(z) = \sum_{k=1}^{\infty} s(k)\frac{(\sigma z)^k(1 - \sigma z)}{1 - z + (1 - \sigma)z(\sigma z)^k}. \tag{2.14}$$

Note that the former expression is not explicit due to the infinite sum. However, using the moment generating property of probability generating functions, we get explicit expressions for the moments of the effective service times for the RAI operation mode.

In particular, the mean value μ_T, the variance ${\sigma_T}^2$ and the skewness γ_T are given by

$$\mu_T = \frac{S\left(\frac{1}{\sigma}\right) - 1}{1-\sigma}, \tag{2.15}$$

$${\sigma_T}^2 = \frac{-\sigma^2 - \sigma(1-\sigma)S\left(\frac{1}{\sigma}\right) + 2\,S\left(\frac{1}{\sigma^2}\right)\sigma - 2\,(1-\sigma)S'\left(\frac{1}{\sigma}\right) - \sigma S\left(\frac{1}{\sigma}\right)^2}{\sigma(1-\sigma)^2} \tag{2.16}$$

and

$$\gamma_T = \frac{\left\{\begin{array}{l} 3\,(1-\sigma)^2 S''\left(\dfrac{1}{\sigma}\right) - 12\,(1-\sigma)S'\left(\dfrac{1}{\sigma^2}\right) \\ + 6\,\sigma(1-\sigma)^2 S'\left(\dfrac{1}{\sigma}\right) + 6\,\sigma(1-\sigma)S'\left(\dfrac{1}{\sigma}\right) S\left(\dfrac{1}{\sigma}\right) \\ - 6\,\sigma^2(1-\sigma)S\left(\dfrac{1}{\sigma^2}\right) - 6\,\sigma^2 S\left(\dfrac{1}{\sigma^2}\right) S\left(\dfrac{1}{\sigma}\right) \\ + 3\,\sigma^2(1-\sigma)S\left(\dfrac{1}{\sigma}\right)^2 + \sigma^2(1-\sigma)^2 S\left(\dfrac{1}{\sigma}\right) \\ - \sigma^3(1+\sigma) + 6\,\sigma^2 S\left(\dfrac{1}{\sigma^3}\right) + 2\,\sigma^2 S\left(\dfrac{1}{\sigma}\right)^3 \end{array}\right\}}{\sigma^2(1-\sigma)^3 {\sigma_T}^3} \tag{2.17}$$

respectively.

In general, the k-th moment is a function of the vacation parameter σ and the probability generating function $S(z)$ and its derivatives evaluated in $z = 1/\sigma^i$ ($i = 1 \ldots k$). This implies that all moments depend on the complete customer service time distribution, or equivalently, on all moments of the customer service times. One then observes that the k-th moment of the effective service time is finite for $\sigma > 1/\sqrt[k]{R_S}$ and infinite for $\sigma < 1/\sqrt[k]{R_S}$. For $\sigma = 1/\sqrt[k]{R_S}$, existence of the k-th moment depends on the convergence of $S(z)$ and its derivatives for $z = R_S$. Here, R_S denotes the common radius of convergence of the probability generating function $S(z)$ and its derivatives.

In many practical cases, the customer service time has an upper bound, say N. In this case, $S(z)$ is polynomial of degree N, the radius of convergence is infinite and (2.14) is an explicit expression as the summation is only over a finite number of values. Now, consider finite radii of convergence. We further assume $R_S > 1$ which in particular implies the existence of all moments of the customer service times. Given $0 < \sigma < 1$, only a finite number of moments of the effective service time are finite. In particular, for a given distribution of the service time, for sufficiently small values of σ, both mean and variance of the effective service time are infinite. When σ increases, the mean becomes finite while the variance remains infinite, meaning that the distribution has a heavy tail. When σ further increases, this tail becomes less and less heavy as

higher moments become finite. For $\sigma = 1$, the effective service time distribution equals the customer service time distribution and all moments are finite due to the assumptions regarding the radius of convergence of $S(z)$. Note that for $R_S = 1$, all moments of the effective service time are infinite for $\sigma < 1$ and equal to the moments of the service time for $\sigma = 1$.

As for CAI, alternative approaches to obtain the probability generating function of the effective service times for RAI are presented in [Fiems et al., 2001] and [Fiems et al., 2002c]. In both contributions, we first obtain the probability generating function of a customer's effective service time conditioned on this customer's service time. Summation over all possible customer service times with respect to the customer service time distribution then yields the unconditional probability generating function of the effective service times. In Fiems et al. [2001], we note that – given the customer service time $S = k$ – the effective service time equals $n > k$ slots if the server is available during the last k slots, is unavailable during the slot preceding these slots, and is not available for more then $k - 1$ slots during the first $n - k - 1$ slots. In Fiems et al. [2002c], we exploit the fact that – given some fixed service time – RAI and RAI,wr operate equivalently.

Repeat after interruption with resampling

The last operation mode under consideration is RAI,wr. The effective service time equals the service time of the customer if no interruption occurs or equals the sum of the length of an unsuccessful attempt and the remaining effective service time after this attempt if this is not the case. That is,

$$T = \begin{cases} S & \text{no interruptions,} \\ \Phi + \tilde{T} & \text{unsuccessful attempt.} \end{cases} \tag{2.18}$$

Here T and $\tilde{T}$ denote the effective service time and the remaining effective service time after an unsuccessful attempt of the customer respectively. S denotes the original service time (the first service time sample) of this customer and Φ denotes the length of an unsuccessful service attempt. The effective service time T and the remaining effective service time $\tilde{T}$ share a common distribution as service is resampled after the unsuccessful attempt. Let $T(z)$ denote the common probability generating function of T and $\tilde{T}$. Conditioning on the length of the original service time then yields,

$$T(z) = \sum_{k=1}^{\infty} s(k) \left(\sigma^k z^k + (1 - \sigma^k) \Phi(z|k) T(z) \right), \tag{2.19}$$

where $\Phi(z|k)$ denotes the probability generating function of the length of an unsuccessful attempt given that the service time equals k slots. The latter probability generating function was already derived during the analysis of RAI. Substitution of this

expression (2.12) in the former then yields

$$T(z) = \frac{S(\sigma z)(1-\sigma z)}{1-z+(1-\sigma)zS(\sigma z)}. \tag{2.20}$$

Again, the moment generating property allows to obtain the various moments of the effective service time for the RAI,wr operation mode. In particular, the mean value μ_T, the variance ${\sigma_T}^2$ and the skewness γ_T are given by,

$$\mu_T = \frac{1-S(\sigma)}{(1-\sigma)S(\sigma)}, \tag{2.21}$$

$${\sigma_T}^2 = \frac{-2\,\sigma(1-\sigma)S'(\sigma)+(1-S(\sigma))(1+\sigma S(\sigma))}{S(\sigma)^2(1-\sigma)^2}, \tag{2.22}$$

and by,

$$\gamma_T = \frac{\left\{\begin{array}{l} -\,3\,\sigma^2 S(\sigma)(1-\sigma)^2 S''(\sigma) \\ +\,6\,\sigma^2(1-\sigma)^2 S'(\sigma)^2 - 6\,\sigma(1-\sigma)S'(\sigma) \\ +\,(1-S(\sigma))(S(\sigma)^2\sigma(1+\sigma)-(1-3\,\sigma)S(\sigma)+2) \end{array}\right\}}{S(\sigma)^3(1-\sigma)^3{\sigma_T}^3}, \tag{2.23}$$

respectively.

In general, the k-th moment of the effective service time for RAI,wr is function of the vacation parameter σ and the probability generating function $S(z)$ and its derivatives evaluated at $z = \sigma$. This implies – as for RAI – that all moments of the effective service times depend on the complete service time distribution. As opposed to RAI however, these moments always exist for $0 < \sigma < 1$, even when the moments of the underlying service time do not exist. For $\sigma = 1$, effective service time and service time are equal as there are no vacations and therefore moments of the effective service time exist whenever the corresponding service time moments exist. The fact that existence conditions are relaxed by the presence of vacations can be explained by noting that excessively long service times will be interrupted and resampled.

An alternative analysis for RAI,wr is presented in our contribution [Fiems et al., 2002c]. There, we retrieve a recursive equation for the probability mass function of the effective service times by conditioning on the state of the server during the first slot of a customer's effective service time. As for CAI, we will frequently follow this approach in the coming sections.

2.1.3 Queue content and customer delay

As the effective service times of the consecutive customers constitute a series of i.i.d. random variables due to the Bernoulli nature of the vacation process, the system can be regarded as a system without vacations, and with customer service times given by the effective service times. I.e., we reduced the analysis of the system under consideration to an equivalent $Geo^X/G/1$ system without vacations as evaluated in section 1.2. Therefore, the queueing system under consideration reaches steady state whenever the effective load $\tilde{\rho}$ is less than 1, i.e.,

$$\tilde{\rho} \triangleq \mu_E \mu_T < 1. \tag{2.24}$$

Recall that μ_E denotes the mean number of arrivals in a slot whereas the mean effective service time μ_T is given by (2.6), (2.15) or (2.21) for the CAI, the RAI or the RAI,wr operation mode respectively. The probability generating functions of the queue content at random slot boundaries and of the customer delay in steady state are then given by,

$$V_r(z) = (1-\tilde{\rho})\frac{(z-1)T(E(z))}{z-T(E(z))}, \tag{2.25}$$

and,

$$D(z) = \frac{1-\tilde{\rho}}{\mu_E}\frac{(z-1)T(z)}{T(z)-1}\frac{E(T(z))-1}{z-E(T(z))}, \tag{2.26}$$

respectively. Here $T(z)$ is given by (2.5), (2.14) and (2.20) for CAI, RAI and RAI,wr respectively.

The moment-generating property of probability generating functions allows us to obtain various moments of these random variables. The reader is referred to section 1.2 for explicit expressions of mean and variance of both queue content at random slot boundaries and customer delay. The mean (variance) of these random variables is a function of mean and variance (mean, variance and skewness) of the effective service times. Expressions for these moments were derived in the preceding subsection for the CAI, RAI and RAI,wr operation modes. These moments can be retrieved from equations (2.6) to (2.8), (2.15) to (2.17) and (2.21) to (2.23) for the CAI, the RAI and the RAI,wr operation modes respectively.

Recall that the system under consideration is weakly stable whenever it reaches steady state, that is, whenever (2.24) is satisfied. The system is strongly stable if the steady-state distribution exists and possesses a finite first moment. As for the system without vacations, it can be seen that the system is strongly stable if the system is weakly stable and the mean values and the variances of both the number of arriving customers in a slot (μ_E, ${\sigma_E}^2$) and the effective service times (μ_T, ${\sigma_T}^2$) are finite. We derived

existence conditions for the latter moments in subsection 2.1.2 for all operation modes under consideration.

2.1.4 Some numerical results

Extensive numerical examples are delayed until section 2.5. For the moment, we limit ourselves to a numerical comparison of the different operation modes under consideration.

Let the efficiency of the operation mode be defined as the following fraction:

$$\epsilon = \frac{\mu_S}{\sigma \mu_T}. \tag{2.27}$$

That is, the efficiency is defined as the mean effective service time when no service is lost divided by the mean effective service time of the operation mode under consideration.

One immediately notes that for CAI operation, no service is lost and therefore the efficiency of this mode always equals 100%. For RAI, the efficiency equals the fraction of available slots during a customer's effective service time where the customer is really served. I.e., lost service slots are not included. Further, one observes that the efficiency depends on the underlying service time distribution and on the server availability probability σ.

Figures 2.4 and 2.5 depict the efficiency for RAI and RAI,wr versus the mean A-period for different service time distributions. In particular, we consider a deterministic, a shifted Poisson and a shifted geometric distribution. The reader may retrieve expressions for the probability generating functions corresponding to these distributions in appendix A. All these distributions are completely characterised by their means. Mean service time equals 5 slots for figure 2.4 and 10 slots for figure 2.5.

In case of RAI,wr, we note that the efficiency equals 100% in case the service time distribution is geometric. Due to the lack of memory of the geometric distribution, CAI and RAI,wr are equivalent as was stated previously. Therefore, efficiency equals 100%. One should further note that in case of RAI,wr efficiency may well be more than 100%. That is, the original service time may be resampled to a shorter one, thereby overcompensating the lost service time.

It can be seen that RAI,wr is always at least as efficient as RAI given the same mean service time μ_S and server availability probability σ. Further, efficiency for RAI and RAI,wr is equal in case of deterministic service times. In this case, RAI and RAI,wr are equivalent as resampled service times always equal the original service time. Given an integer mean service time μ_S, one obtains the best efficiency for RAI

and the worst efficiency for RAI,wr if the probability generating function of the service times is deterministic.

These observations also follow from Jensen's inequality in a general context. Jensen's inequality states that for any function $u(x)$, convex in some open interval, and a random variable X that only takes values within this interval, we have,

$$\mathrm{E}[u(X)] \geq u(\mathrm{E}[X]). \tag{2.28}$$

A function $u(x)$ is convex over an open interval (a, b), if following inequality holds for every $x_1, x_2 \in (a, b)$ and for every $t \in [0, 1]$,

$$u(tx_1 + (1-t)x_2) \leq tu(x_1) + (1-t)u(x_2). \tag{2.29}$$

In particular, for $\phi > 0$, the function $u(x) = \phi^x$ is convex for all $x > 0$. Customer service times are positive. Therefore equation (2.28) leads to,

$$S(\phi) = \mathrm{E}\left[\phi^S\right] \geq \phi^{\mathrm{E}[S]} = \phi^{\mu_S}. \tag{2.30}$$

Plugging $\phi = \sigma$ and $\phi = 1/\sigma$ into the former inequality then yields:

$$S\left(\frac{1}{\sigma}\right) \geq \frac{1}{\sigma^{\mu_S}} \geq \frac{1}{\hat{S}(\sigma)}. \tag{2.31}$$

Here $S(z)$ and $\hat{S}(z)$ denote two different probability generating functions of customer service times with equal mean μ_S. In view of the expressions (2.15) and (2.21) for the mean effective service times for RAI and RAI,wr respectively, the former inequalities now easily lead to the conclusion that the efficiency for RAI,wr is always as good as or better than the efficiency for RAI given the mean customer service time and given the server availability probability σ. Further, given an integer mean customer service time and given the server availability probability σ, the deterministic distribution – in comparison with all possible customer service time distributions – yields the worst efficiency for RAI,wr and the best efficiency for RAI. The latter observations follow from the fact that RAI and RAI,wr operate equivalently for deterministically distributed customer service times and that the efficiency for RAI,wr is at least as good as the efficiency for RAI given the mean customer service time and given the server availability probability.

2.2 The method of the supplementary variables

Consider a queueing system at a random slot boundary. In most cases, future queueing behaviour depends on the past, given the queue content at that particular epoch. How-

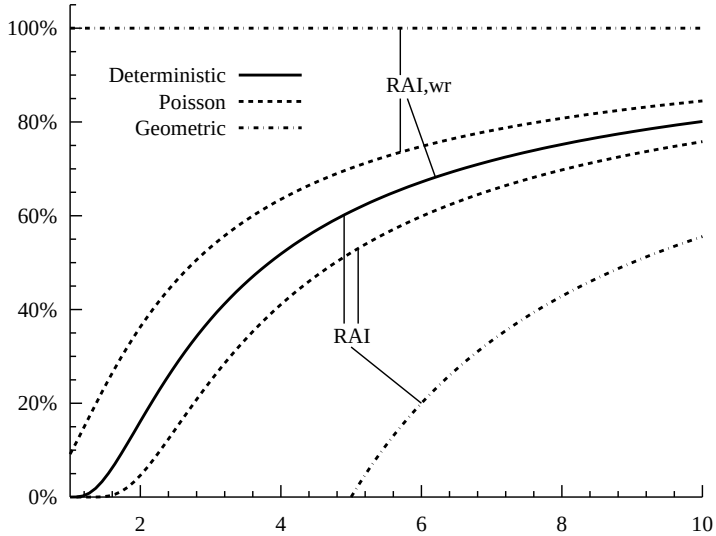


Figure 2.4: Efficiency ϵ vs. the mean A-period μ_A for deterministically, Poisson en geometrically distributed customer service times. (The mean customer service time μ_S equals 5 slots.)

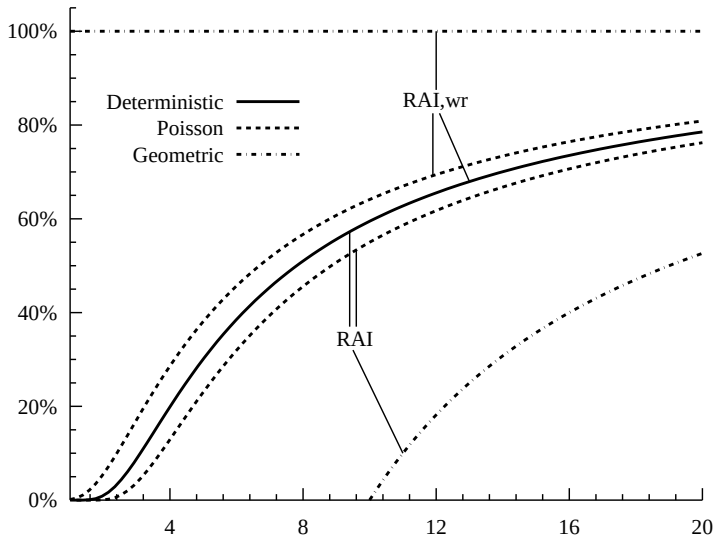


Figure 2.5: Efficiency ϵ vs. the mean A-period μ_A for deterministically, Poisson en geometrically distributed customer service times. (The mean customer service time μ_S equals 10 slots.)

ever, it is often possible to add a limited number of *state variables*, such that, given these state variables, future queueing behaviour is independent of the past. I.e., queue content and state variables provide a system state description in the Markovian sense. This is the so-called method of the supplementary variables.

This method is ascribed – see e.g., Stidham [2002] or Takagi [1991] – to Cox [1955]. Different authors have applied the method to analyse discrete-time queueing systems. For example, Bruneel [1993] investigates the discrete-time $Geo^X/G/1$ queue whereas both Takahashi et al. [1999] and Lee [2001] consider variants of this system. The former investigates a retrial queue with non-preemptive priorities whereas the latter considers a preemptive resume priority queueing system.

In the following subsections, we present the supplementary variable analysis of the Bernoulli interrupted $Geo^X/G/1$ queue for CAI, RAI,wr and RAI operation modes as an alternative to the effective service time methodology developed in the previous section. We show how to retrieve various performance measures and conclude the section with a comparison of both methodologies.

2.2.1 Continue after interruption

In case of the CAI operation mode, at the beginning of any slot, say slot k, the state of the system is completely described by the number of customers present in the system $V_r^{(k)}$ and by the remaining customer service time $H^{(k)}$ of the customer in service (if any). If there is no customer in service, we define that $H^{(k)} = 0$. That is, $V_r^{(k)} = 0$ implies $H^{(k)} = 0$. As the remaining service time of a customer in service equals at least 1 slot, one may note that $H^{(k)} = 0$ also implies $V_r^{(k)} = 0$.

We can now relate the system state $(V_r^{(k+1)}, H^{(k+1)})$ at the beginning of slot $k+1$ to the system state $(V_r^{(k)}, H^{(k)})$ at the beginning of slot k by the following set of system equations.

- If $(H^{(k)} = V_r^{(k)} = E^{(k)} = 0) \vee (H^{(k)} = V_r^{(k)} = Q^{(k)} = 1 \wedge E^{(k)} = 0)$, then

$$\begin{aligned} V_r^{(k+1)} &= 0, \\ H^{(k+1)} &= 0. \end{aligned} \tag{2.32}$$

  I.e., if the system is empty and there are no new arrivals, or if the last customer leaves the system and there are no new arrivals, then the system is empty at the beginning of the next slot.

- If $(H^{(k)} > 0 \wedge Q^{(k)} = 0) \vee (H^{(k)} > 1 \wedge Q^{(k)} = 1)$, then,

$$\begin{aligned} V_r^{(k+1)} &= V_r^{(k)} + E^{(k)}, \\ H^{(k+1)} &= H^{(k)} - Q^{(k)}. \end{aligned} \tag{2.33}$$

  That is, given that there's a customer in service and that this customer will not leave the system at the end of slot k, the arriving customers are queued.

- If $(H^{(k)} = V_r^{(k)} = 0 \wedge E^{(k)} > 0) \vee (H^{(k)} = Q^{(k)} = 1 \wedge V_r^{(k)} > 1)$ $\vee (H^{(k)} = V_r^{(k)} = Q^{(k)} = 1 \wedge E^{(k)} > 0)$, then

$$\begin{aligned} V_r^{(k+1)} &= (V_r^{(k)} - 1)^+ + E^{(k)}, \\ H^{(k+1)} &= S^{(k)}. \end{aligned} \tag{2.34}$$

  Finally, a new customer starts service, if the system is or becomes empty and there are new arrivals , or if a customer leaves a non-empty system.

As before, $E^{(k)}$ denotes the number of customer arrivals during slot k, $Q^{(k)}$ denotes the number of available servers during slot k and $S^{(k)}$ denotes the service time of the customer that starts service just after slot k. We use $\wedge$ and $\vee$ to denote the logical *and* and *or* operators respectively.

The former equations then allow to relate the joint probability generating functions of remaining service time and system content at the beginning of slots k and $k+1$. Some standard z-transform manipulations yield

$$\begin{aligned} P^{(k+1)}(x,z) &\triangleq \mathrm{E}\left[x^{H^{(k+1)}} z^{V_r^{(k+1)}}\right] \\ &= (E(0) + (E(z) - E(0))S(x))\, P^{(k)}(0,0) \\ &+ (E(0) + (E(z) - E(0))S(x))\, \sigma\Omega^{(k)}(0) \\ &+ (1-\sigma)E(z)\left(P^{(k)}(x,z) - P^{(k)}(0,0)\right) \\ &+ \sigma S(x)E(z)\left(\Omega^{(k)}(z) - \Omega^{(k)}(0)\right) \\ &+ \sigma\frac{E(z)}{x}\left(P^{(k)}(x,z) - xz\Omega^{(k)}(z) - P^{(k)}(0,0)\right), \end{aligned} \tag{2.35}$$

with,

$$\Omega^{(k)}(z) \triangleq \mathrm{E}\left[z^{V_r^{(k)}-1} | H^{(k)} = 1\right] \Pr\left[H^{(k)} = 1\right]. \tag{2.36}$$

Under the assumption that the system under consideration reaches steady state, let $P(x,z) = \lim_{k\to\infty} P^{(k)}(x,z)$ denote the steady-state joint probability generating

function. The former expressions then yield,

$$P(x,z) = \frac{\left\{\begin{array}{l} \sigma x E(0)\Omega(0)\left(1-S(x)\right) + \sigma x E(z)\Omega(z)\left(S(x)-z\right) \\ +P(0,0)x\left(E(0)+(E(z)-E(0))S(x)\right) \\ -P(0,0)E(z)\left((1-\sigma)x+\sigma\right) \end{array}\right\}}{x-(1-\sigma)E(z)x-\sigma E(z)}, \tag{2.37}$$

with $\Omega(z) = \lim_{k\to\infty}\Omega^{(k)}(z)$. As $P(x,z)$ is a joint probability generating function, it is bounded for all $|x|, |z| \leq 1$. In particular, let x equal

$$\Psi(z) = \frac{\sigma E(z)}{1-(1-\sigma)E(z)}. \tag{2.38}$$

One easily verifies that the former expression is a probability generating function and therefore $|\Psi(z)| \leq 1$ for $z \leq 1$. This leads to the conclusion that the numerator in (2.37) vanishes for values $(\Psi(z), z)$ for $|z| \leq 1$ as the denominator vanishes for these values. This observation yields,

$$\begin{aligned} \Omega(z) = P(0,0)&\frac{E(0)-1+\left(E(z)-E(0)\right)S\left(\Psi(z)\right)}{\sigma E(z)(z-S\left(\Psi(z)\right))} \\ &+\Omega(0)\frac{E(0)}{E(z)}\frac{1-S\left(\Psi(z)\right)}{z-S\left(\Psi(z)\right)}. \end{aligned} \tag{2.39}$$

Substitution of $z=0$ in the former equation then allows us to determine $\Omega(0)$,

$$\Omega(0) = P(0,0)\frac{1-E(0)}{\sigma E(0)}. \tag{2.40}$$

which further simplifies equation (2.39) to

$$\Omega(z) = P(0,0)\frac{E(z)-1}{\sigma E(z)}\frac{S\left(\Psi(z)\right)}{z-S\left(\Psi(z)\right)}. \tag{2.41}$$

Plugging the former two equations into (2.37), yields the following expression for the joint probability generating function $P(x,z)$,

$$P(x,z) = P(0,0)\frac{\left\{\begin{array}{l} x(1-E(0))(1-S(x))(z-S\left(\Psi(z)\right)) \\ +x(1-E(z))S\left(\Psi(z)\right)\left(z-S(x)\right) \\ +(E(z)-E(0))S(x)x(z-S\left(\Psi(z)\right)) \\ +(E(0)x-(\sigma+(1-\sigma)x)E(z))(z-S\left(\Psi(z)\right)) \end{array}\right\}}{\left(x-(1-\sigma)E(z)x-\sigma E(z)\right)\left(z-S\left(\Psi(z)\right)\right)}. \tag{2.42}$$

Here, only the constant $P(0,0)$ remains unknown. By means of the normalisation condition $P(1,1) = 1$, we can retrieve the latter as well,

$$P(0,0) = 1 - \frac{\mu_S \mu_E}{\sigma}. \tag{2.43}$$

The expression (2.42) of the joint probability generating function allows us to retrieve expressions for a number of performance measures as we will see further on.

2.2.2 Repeat after interruption with resampling

As for CAI, the queue content $V_r^{(k)}$ and the remaining service time $H^{(k)}$ provide a sufficient state description at a random slot boundary, say slot k, in case of the RAI,wr operation mode. Again, we assume that $H^{(k)} = 0$ if the system is empty. As for CAI, one easily sees that $V_r^{(k)}$ and $H^{(k)}$ or either both non-zero or both zero.

The state variables at the beginning of slot $k+1$ relate to those at the beginning of slot k as follows.

- If $(H^{(k)} = V_r^{(k)} = E^{(k)} = 0) \vee (H^{(k)} = V_r^{(k)} = Q^{(k)} = 1 \wedge E^{(k)} = 0)$, then

$$\begin{aligned} V_r^{(k+1)} &= 0, \\ H^{(k+1)} &= 0. \end{aligned} \tag{2.44}$$

  I.e., the system is empty at the beginning of slot $k+1$ if there are no arrivals during slot k and if either there are no customers in the system at the beginning of slot k or the last customer leaves the system at the end of slot k.

- If $H^{(k)} > 1 \wedge Q^{(k)} = 1$, then,

$$\begin{aligned} V_r^{(k+1)} &= V_r^{(k)} + E^{(k)}, \\ H^{(k+1)} &= H^{(k)} - 1. \end{aligned} \tag{2.45}$$

  That is, given that the customer in service does not finish his service during slot k and given that the server is available, the remaining service time is diminished by one and possible arrivals are queued.

- If $(H^{(k)} = V_r^{(k)} = 0 \wedge E^{(k)} > 0) \vee (H^{(k)} = Q^{(k)} = 1 \wedge V_r^{(k)} > 1)$ $\vee (V_r^{(k)} = H^{(k)} = Q^{(k)} = 1 \wedge E^{(k)} > 0) \vee (H^{(k)} > 0 \wedge Q^{(k)} = 0)$, then

$$\begin{aligned} V_r^{(k+1)} &= \left(V_r^{(k)} - Q^{(k)}\right)^+ + E^{(k)}, \\ H^{(k+1)} &= S^{(k)}. \end{aligned} \tag{2.46}$$

That is, in all other cases, service (re)starts with a new service time sample.

As before, $E^{(k)}$ denotes the number of customer arrivals during slot k and $Q^{(k)}$ denotes the number of available servers during slot k. Further, $S^{(k)}$ denotes the service time of the customer that starts or restarts service just after slot k.

The former set of system equations then translate into a corresponding relation between the joint probability generating functions $P^{(k)}(x,z)$ and $P^{(k+1)}(x,z)$ of remaining service time and queue content at the beginning of slots k and $k+1$ respectively,

$$\begin{aligned} P^{(k+1)}(x,z) &\triangleq \mathrm{E}\left[x^{H^{(k+1)}} z^{V_r^{(k+1)}}\right] \\ &= (E(0) + (E(z) - E(0))S(x))\, P^{(k)}(0,0) \\ &+ \sigma\,(E(0) + (E(z) - E(0))S(x))\,\Omega^{(k)}(0) \\ &+ (1-\sigma)E(z)S(x)\left(P^{(k)}(1,z) - P^{(k)}(0,0)\right) \\ &+ \sigma S(x)E(z)\left(\Omega^{(k)}(z) - \Omega^{(k)}(0)\right) \\ &+ \sigma\frac{E(z)}{x}\left(P^{(k)}(x,z) - xz\Omega^{(k)}(z) - P^{(k)}(0,0)\right). \end{aligned} \tag{2.47}$$

Here, the partial probability generating function $\Omega^{(k)}(z)$ is again defined as

$$\Omega^{(k)}(z) = \mathrm{E}\left[z^{V_r^{(k)}-1} | H^{(k)} = 1\right] \Pr\left[H^{(k)} = 1\right]. \tag{2.48}$$

Under the assumption that the system under consideration reaches steady state, let $P(x,z) = \lim_{k\to\infty} P^{(k)}(x,z)$ denote the steady-state joint probability generating function of the remaining service time and the queue content in steady state. From equation (2.47) we easily get

$$P(x,z) = \frac{\left\{\begin{array}{l} x\,(E(0) + (E(z) - E(0))S(x))\,(P(0,0) + \sigma\Omega(0)) \\ + (1-\sigma)xE(z)S(x)\,(P(1,z) - P(0,0)) \\ + \sigma x S(x)E(z)\,(\Omega(z) - \Omega(0)) \\ - \sigma E(z)\,(xz\Omega(z) + P(0,0)) \end{array}\right\}}{x - \sigma E(z)}, \tag{2.49}$$

with $\Omega(z) = \lim_{k\to\infty} \Omega^{(k)}(z)$ the steady-state partial probability generating function corresponding to $\Omega^{(k)}(z)$. After substitution of $x = 1$ in the former expression, we can solve for $P(1,z)$,

$$P(1,z) = \Omega(z)\sigma E(z)\frac{1-z}{1-E(z)}. \tag{2.50}$$

As $P(x,z)$ is a joint probability generating function, it is bounded for all $|x|, |z| \leq 1$. In particular, let x equal $\sigma E(z)$. One easily verifies that $|\sigma E(z)| \leq 1$ for all $|z| \leq 1$. Therefore, as the denominator in (2.49) disappears for $(\sigma E(z), z)$ for $|z| \leq 1$, the numerator has to disappear as well. This observation then leads to the following expression for the unknown function $\Omega(z)$,

$$\Omega(z) = \frac{\left\{\begin{array}{l} P(0,0)\left(1 - E(0) - (\sigma E(z) - E(0))S(\sigma E(z))\right) \\ - \Omega(0)\sigma E(0)(1 - S(\sigma E(z))) \\ - P(1,z)(1-\sigma)E(z)S(\sigma E(z)) \end{array}\right\}}{\sigma E(z)(S(\sigma E(z)) - z)}. \tag{2.51}$$

Substitution of expression (2.50) in the former expression and evaluation in $z = 0$ then yields,

$$\Omega(0) = P(0,0)\frac{1 - E(0)}{\sigma E(0)}. \tag{2.52}$$

We can now plug the former expression and equation (2.50) into equation (2.51). The latter then simplifies to,

$$\Omega(z) = \frac{P(0,0)S(\sigma\, E(z))(1 - E(z))(1 - \sigma E(z))}{\sigma E(z)[(1 - (\sigma + (1-\sigma)z)E(z))S(\sigma E(z)) - z(1 - E(z))]}. \tag{2.53}$$

By substitution of equations (2.50), (2.52) and (2.53) into (2.49) , we obtain the following expression for the joint probability generating function $P(x,z)$ where only the constant $P(0,0)$ remains unknown,

$$\begin{aligned} P(x,z) =& \frac{P(0,0)}{x - \sigma E(z)} \\ & \times \frac{\left\{\begin{array}{l} \left((\sigma(1-z) + z(1-x))\, E(z) + xz - 1\right)\sigma E(z)S(\sigma E(z)) \\ + x(1-z)(1 - \sigma E(z))S(\sigma E(z)) \\ - (1 - E(z))\, z\, (\sigma S(x)E(z)x - \sigma E(z) + x - xS(x)) \end{array}\right\}}{(1 - (\sigma + (1-\sigma)z)E(z))S(\sigma E(z)) - z\,(1 - E(z))}. \end{aligned} \tag{2.54}$$

The latter constant then follows from the normalisation condition $P(1,1) = 1$ of the joint probability generating function,

$$P(0,0) = 1 - \mu_E \frac{1 - S(\sigma)}{(1-\sigma)S(\sigma)}. \tag{2.55}$$

As for CAI, we may retrieve a number of performance measures from the former results as we will see further.

2.2.3 Repeat after interruption

In the case of the repeat after interruption operation mode, the queue content $V_r^{(k)}$ and the remaining service time $H^{(k)}$ no longer provide a sufficient state description at a random slot boundary, say slot k. Additionally, one needs to keep track of the customer service time of the customer in service $G^{(k)}$ as the latter starts over in case of an interruption. We assume that $H^{(k)} = G^{(k)} = 0$ if there is no customer in service. As $G^{(k)}$ and $H^{(k)}$ take positive values whenever a customer is in service, one sees that the state variables are either all non-zero or all equal to zero.

We again relate the state variables at the beginning of slot $k+1$ to those at the beginning of slot k. We get following set of system equations.

- If $(H^{(k)} = E^{(k)} = 0) \vee (H^{(k)} = V_r^{(k)} = Q^{(k)} = 1 \wedge E^{(k)} = 0)$, than,

$$\begin{aligned} V_r^{(k+1)} &= 0. \\ H^{(k+1)} &= 0, \\ G^{(k+1)} &= 0. \end{aligned} \tag{2.56}$$

  That is, given that the queue either is or becomes empty and given that there are no new arrivals, the queue is empty at the beginning of the next slot.

- If $(H^{(k)} = Q^{(k)} = 1 \wedge V_r^{(k)} > 1) \vee (H^{(k)} = 0 \wedge E^{(k)} > 0)$ $\vee (H^{(k)} = Q^{(k)} = V_r^{(k)} = 1 \wedge E^{(k)} > 0)$, then,

$$\begin{aligned} V_r^{(k+1)} &= (V_r^{(k)} - 1)^+ + E^{(k)}, \\ H^{(k+1)} &= S^{(k)}, \\ G^{(k+1)} &= S^{(k)}. \end{aligned} \tag{2.57}$$

  That is, given that a customer leaves the queue and that the latter is not empty at the beginning of slot $k+1$, or given that a new customer arrives in an empty queue, a new customer starts service.

- If $Q^{(k)} = 1 \wedge H^{(k)} > 1$, then,

$$\begin{aligned} V_r^{(k+1)} &= V_r^{(k)} + E^{(k)}, \\ H^{(k+1)} &= H^{(k)} - 1, \\ G^{(k+1)} &= G^{(k)}. \end{aligned} \tag{2.58}$$

  I.e., given that the server is available and that the customer in service needs more slots of service, the latter continues service.

- If $Q^{(k)} = 0 \wedge H^{(k)} > 0$, then,

$$\begin{aligned} V_r^{(k+1)} &= V_r^{(k)} + E^{(k)}, H^{(k+1)} &= G^{(k)}, \\ G^{(k+1)} &= G^{(k)}. \end{aligned} \tag{2.59}$$

  Finally, given that the server is unavailable and that a customer is in service, service of this customer has to start over.

The former system equations translate in a relation between the joint probability generating functions $\tilde{P}^{(k)}(x,y,z)$ and $\tilde{P}^{(k+1)}(x,y,z)$ of the service time of the customer in service, of the remaining service time of this customer and of the queue content at the beginning of slots k and $k+1$ respectively,

$$\begin{aligned} \tilde{P}^{(k+1)}(x,y,z) &\triangleq \mathrm{E}\left[x^{H^{(k)}} y^{G^{(k)}} z^{V_r^{(k)}}\right] \\ &= \left(E(0) + S(xy)(E(z) - E(0))\right) \tilde{P}^{(k)}(0,0,0) \\ &+ \sigma \left(E(0) + S(xy)(E(z) - E(0))\right) \Theta^{(k)}(1,0) \\ &+ \sigma S(xy) E(z) \left(\Theta^{(k)}(1,z) - \Theta^{(k)}(1,0)\right) \\ &+ \sigma \frac{E(z)}{x} \left(\tilde{P}^{(k)}(x,y,z) - xz\Theta^{(k)}(y,z) - \tilde{P}^{(k)}(0,0,0)\right) \\ &+ (1-\sigma) E(z) \left(\tilde{P}^{(k)}(1,xy,z) - \tilde{P}^{(k)}(0,0,0)\right). \end{aligned} \tag{2.60}$$

Here, the unknown function $\Theta^{(k)}(y,z)$ is defined as the following partial joint probability generating function,

$$\Theta^{(k)}(y,z) = \mathrm{E}\left[y^{G^{(k)}} z^{V_r^{(k)}-1} | H^{(k)} = 1\right] \Pr\left[H^{(k)} = 1\right]. \tag{2.61}$$

Under the assumption that the system reaches steady state, we denote the steady-state joint probability generating function of the remaining service time of the customer in service, of the service time of the customer in service and of the queue content by $\tilde{P}(x,y,z) = \lim_{k\to\infty} \tilde{P}^{(k)}(x,y,z)$. From the former expressions we get,

$$\tilde{P}(x,y,z) = \frac{\left\{\begin{array}{l} x\left(E(0) + S(xy)(E(z) - E(0))\right) \tilde{P}(0,0,0) \\ -(\sigma + (1-\sigma)x) E(z) \tilde{P}(0,0,0) \\ + \sigma x E(0)(1 - S(xy))\Theta(1,0) - \sigma E(z) xz \Theta(y,z) \\ + \sigma x S(xy) E(z) \Theta(1,z) + (1-\sigma) x E(z) \tilde{P}(1,xy,z) \end{array}\right\}}{x - \sigma E(z)}, \tag{2.62}$$

with $\Theta(y,z) = \lim_{k\to\infty} \Theta^{(k)}(y,z)$. Substitution of $x = 1$ in the former equation,

allows us to solve for $\tilde{P}(1, y, z)$,

$$\tilde{P}(1, y, z) = \frac{\left\{\begin{array}{l} (S(y) - 1)\,(E(z) - E(0))\,\tilde{P}(0, 0, 0) \\ + \sigma E(z) S(y) \Theta(1, z) - \sigma E(z) z \Theta(y, z) \\ - \sigma E(0)\,(S(y) - 1)\,\Theta(1, 0) \end{array}\right\}}{1 - E(z)}. \tag{2.63}$$

The former expression enables us to remove the unknown probability generating function $\tilde{P}(1, xy, z)$ from the right-hand side of equation (2.62),

$$\tilde{P}(x, y, z) = \frac{\left\{\begin{array}{l} xS(xy)(1 - \sigma E(z))(E(z) - E(0))\tilde{P}(0, 0, 0) \\ + xS(xy)(1 - \sigma E(z))\sigma E(z)\Theta(1, z) \\ - \sigma xzE(z)((1 - \sigma)E(z)\Theta(xy, z) + (1 - E(z))\Theta(y, z)) \\ + \sigma E(0)(1 - S(xy))(1 - \sigma E(z))x\Theta(1, 0) \\ + ((1 - \sigma E(z))E(0)x + \sigma E(z)^2)\tilde{P}(0, 0, 0) \\ - ((1 - \sigma)x + \sigma)E(z)\tilde{P}(0, 0, 0) \end{array}\right\}}{(1 - E(z))(x - \sigma E(z))}. \tag{2.64}$$

Similarly as for CAI and RAI,wr, the joint probability generating function $\tilde{P}(x, y, z)$ is bounded for $|x|, |y|, |z| \leq 1$ implying that wherever the denominator of the right-hand side of (2.64) disappears, the numerator must disappear as well. In particular, the numerator must disappear for $(x, y, z) = (\sigma E(0), 1, 0)$, which yields,

$$\Theta(1, 0) = \tilde{P}(0, 0, 0)\frac{1 - E(0)}{\sigma E(0)}. \tag{2.65}$$

Further, the denominator also disappears for all $(x, y, z) = (\sigma E(z), y, z)$ for $|y|, |z| \leq 1$. As the corresponding numerator disappears as well for this 3-tuple, we find the following expression for the unknown function $\Theta(y, z)$:

$$\Theta(y, z) = \frac{\left\{\begin{array}{l} (1 - \sigma)\sigma zE(z)^2\Theta(\sigma yE(z), z) \\ - (1 - \sigma E(z))\sigma E(z)S(\sigma yE(z))\Theta(1, z) \\ + (1 - E(z))(1 - \sigma E(z))S(y\sigma E(z))\tilde{P}(0, 0, 0) \end{array}\right\}}{\sigma E(z)z(E(z) - 1)}. \tag{2.66}$$

Here, we also plugged in expression (2.65) to simplify the result.

The former functional equation for $\Theta(y, z)$ does not allow us to obtain expressions for the various derivatives evaluated at 1. We can however transform the former expression as follows. Let $\Theta(z|i)$ denote the following partial conditional probability

generating function,

$$\Theta(z|i) \triangleq \mathrm{E}\left[z^{V_r-1}|H=1, G=i\right] \Pr\left[H=1\right]. \tag{2.67}$$

One then easily verifies that the latter relates to $\Theta(y,z)$ as,

$$\Theta(y,z) = \sum_{i=1}^{\infty} \Pr[G=i|H=1]\Theta(z|i)y^i = \sum_{i=1}^{\infty} s(i)\Theta(z|i)y^i. \tag{2.68}$$

The last equality here follows from the fact that the total service time of a customer in service during a random departure slot equals the customer service time of a random customer as every customer only departs once. A random departure slot is a slot where the remaining service time of the customer in service equals 1 and where the server is available. That is, we get,

$$s(i) = \Pr[G=i|H=1, Q=1] = \Pr[G=i|H=1], \tag{2.69}$$

for $i \geq 1$. The i.i.d. nature of the vacation process explains the last equality.

Plugging equation (2.68) into (2.66) now yields

$$\sum_{i=1}^{\infty} s(i)\Theta(z|i)\left(1 + \frac{(1-\sigma)E(z)}{1-E(z)}(\sigma E(z))^i\right)y^i = \sum_{i=1}^{\infty} s(i)(\sigma E(z))^i y^i \frac{(1-\sigma E(z))}{z}\left(\frac{\Theta(1,z)}{1-E(z)} - \frac{\tilde{P}(0,0,0)}{\sigma E(z)}\right). \tag{2.70}$$

from which we retrieve an explicit expression for the partial conditional probability generating functions $\Theta(z|i)$ for each $i \geq 1$ by termwise comparison of the coefficients of y^i,

$$\Theta(z|i) = \frac{(\sigma E(z))^i(\sigma E(z)-1)\left[\tilde{P}(0,0,0)(1-E(z)) - \Theta(1,z)\sigma E(z)\right]}{\sigma E(z) z\left[1-E(z)+(1-\sigma)E(z)(\sigma E(z))^i\right]}. \tag{2.71}$$

Substitution of the former expression into equation (2.68) yields

$$\Theta(y,z) = \frac{\Gamma(y,z)}{z}\left(\Theta(1,z) - \tilde{P}(0,0,0)\frac{1-E(z)}{\sigma E(z)}\right), \tag{2.72}$$

where – for ease of notation – we define $\Gamma(y,z)$ as,

$$\Gamma(y,z) \triangleq \sum_{i=1}^{\infty} s(i)\frac{(\sigma E(z))^i(1-\sigma E(z))y^i}{1-E(z)+(1-\sigma)E(z)(\sigma E(z))^i}. \tag{2.73}$$

Plugging $y = 1$ into equation (2.72), allows us to solve for the unknown function $\Theta(1, z)$. This yields

$$\Theta(1, z) = \tilde{P}(0, 0, 0)\frac{(1 - E(z))\Gamma(1, z)}{\sigma E(z)(\Gamma(1, z) - z)}. \tag{2.74}$$

This result then further simplifies equation (2.72) to

$$\Theta(y, z) = \frac{\tilde{P}(0, 0, 0)(1 - E(z))\Gamma(y, z)}{\sigma E(z)(\Gamma(1, z) - z)}. \tag{2.75}$$

One easily verifies that the former expression for $\Theta(y, z)$ satisfies the functional equation (2.66).

Bringing everything together, we finally get

$$\tilde{P}(x, y, z) = \tilde{P}(0, 0, 0)\frac{\left\{\begin{array}{l}(x - \sigma E(z))\Gamma(1, z) - xzE(z)(1 - \sigma)\Gamma(xy, z)\\ - xz\,(1 - E(z))\,\Gamma(y, z) + \sigma E(z)z\\ + xz\,(1 - \sigma E(z))\,S(xy) - xz\end{array}\right\}}{(x - \sigma E(z))(\Gamma(1, z) - z)} \tag{2.76}$$

The only remaining unknown in the former expression is $\tilde{P}(0, 0, 0)$ as $\Gamma(y, z)$ is explicitly displayed in (2.73).

As before, the normalisation condition $\tilde{P}(1, 1, 1) = 1$ then allows us to retrieve the remaining unknown factor $\tilde{P}(0, 0, 0)$,

$$\tilde{P}(0, 0, 0) = 1 - \mu_E\frac{S\left(\frac{1}{\sigma}\right) - 1}{1 - \sigma}. \tag{2.77}$$

Recall that $P(x, z)$ and $\Omega(z)$ denote the joint probability generating function of the remaining service time and queue content and the partial probability generating function of the queue content minus one given that the remaining service time of the customer in service equals 1 slot respectively. In view of the definitions (2.60) and (2.61) of $\tilde{P}(x, y, z)$ and $\Theta(y, z)$, we easily find that these generating functions for RAI are

given by

$$P(x,z) \triangleq \mathrm{E}\left[x^H z^{V_r}\right] \tag{2.78}$$
$$= \tilde{P}(x,1,z)$$
$$= \left(1-\mu_E \frac{S\left(\frac{1}{\sigma}\right)-1}{1-\sigma}\right) \frac{\left\{\begin{array}{l}(x-\sigma E(z)-xz(1-E(z)))\Gamma(1,z)\\ -xzE(z)(1-\sigma)\Gamma(x,z)+\sigma E(z)z\\ +xz\left(1-\sigma E(z)\right)S(x)-xz\end{array}\right\}}{(x-\sigma E(z))(\Gamma(1,z)-z)} \tag{2.79}$$

and

$$\Omega(z) \triangleq \mathrm{E}\left[z^{V_r-1}|H=1\right]\Pr[H=1] \tag{2.80}$$
$$= \Theta(1,z)$$
$$= \left(1-\mu_E \frac{S\left(\frac{1}{\sigma}\right)-1}{1-\sigma}\right) \frac{(1-E(z))\Gamma(1,z)}{\sigma E(z)(\Gamma(1,z)-z)}. \tag{2.81}$$

As for the CAI and the RAI,wr operation modes, we may retrieve a number of performance measures from the former results as we will see further.

2.2.4 Performance measures

We can now distill a number of performance measures from the expressions of the joint probability generating functions that were derived in the preceding subsections.

We can a.o. obtain expressions for the probability generating functions of the queue content at various epochs in time. As before, $V_r(z)$, $V_d(z)$ and $V_s(z)$ denote the probability generating functions of the queue content at random slot boundaries, at departure times and at start of service epochs. We get:

$$V_r(z) = \mathrm{E}\left[z^{V_r}\right]$$
$$= P(1,z), \tag{2.82}$$
$$V_d(z) = \mathrm{E}\left[z^{V_r-1+E}\middle| H=1 \wedge Q=1\right]$$
$$= \frac{\Omega(z)}{\Omega(1)}E(z) \tag{2.83}$$
$$\tag{2.84}$$

and

$$\begin{aligned} V_s(z) &= \mathrm{E}\left[z^{(V_r-1)^+ + E}\middle|(H = Q = 1 \wedge V_r > 1)\vee \right. \\ &\qquad \left. (H = V_r = Q = 1 \wedge E > 0) \vee (H = 0 \wedge E > 0)\right] \\ &= \frac{\sigma E(z)(\Omega(z) - \Omega(0)) + (E(z) - E(0))(\sigma\Omega(0) + P(0,0))}{\sigma(\Omega(1) - \Omega(0)) + (1 - E(0))(\sigma\Omega(0) + P(0,0))}. \end{aligned} \tag{2.85}$$

Here, equation (2.83) follows from the fact that a customer departs from the system at the end of a slot where his remaining service time equals 1 slot and where the server is available. Equation (2.84) follows from the fact that a customer starts service during the slot following a slot where a customer leaves a non-empty system or where an arrival occurs in an empty system or in a system becoming empty. The generating functions $P(x, z)$ and $\Omega(z)$ are given by (2.42) and (2.41), (2.54) and (2.51) and (2.79) and (2.81) for CAI, RAI,wr and RAI operation modes respectively.

Results are not limited to probability generating functions of queue content. For example, let *total unfinished service time* on random slot boundaries denote the number of slots necessary to return to an empty system if one assumes that there are neither new arrivals nor vacations. The total unfinished service time then equals the sum of the remaining service time of the customer in service and the service times of all other customers present in the system. If there are no customers in the system, the total unfinished service time equals 0. One easily verifies that the probability generating function $\tilde{U}(z)$ of the total unfinished service time is given by

$$\tilde{U}(z) = P(0,0) + \frac{P(z, S(z)) - P(0,0)}{S(z)}. \tag{2.86}$$

Note that total unfinished service time does not equal unfinished work. The latter was defined (see section 1.2.3) as the number of slots it takes to return to an empty system under the assumption that there are no new customer arrivals. That is, unfinished work takes server vacations into account. Depending on the application, one may retrieve other results as well. However, it does not seem possible to obtain probability generating functions of unfinished work and customer delay without retrieving generating functions of effective service times first.

2.2.5 Comparison

We are now able to compare the effective service time (EST) and the supplementary variable (SV) approach. A comparison requires us to assess various difficulties encountered during the analysis. It is worth pointing out that such an assessment is always highly subjective.

Compared to the SV approach, the EST approach is more compact because we can use the readily available results of the $Geo^X/G/1$ queueing system and because the queueing analysis itself is unified for all operation modes once we have obtained the probability generating functions of the effective service times. For more complex (correlated) vacation queueing models, the EST approach still allows a unified queueing analysis but knowledge of the probability generating functions of the effective service times no longer reduces the queueing model to a simple $Geo^X/G/1$ queueing system as we shall see further.

The EST approach is an ad-hoc approach, tailored for the analysis of some particular queueing models with vacations. The range of problems that can be solved by this approach is fairly limited and some degree of stochastic insight is required during analysis. On the other hand, the SV approach yields results for a broad class of queueing problems. The approach offers a set of standard techniques which translate the problem formulation (the system equations) in a fairly straightforward manner into the corresponding probability generating function results. Although involved formulas are often lengthy, this does not add a lot of difficulties as one may rely an algebraic mathematical software for formula manipulation.

Regarding results, both methods are complementary. That is, although some performance measures may be retrieved from either method, there are quite a number of results that can only be (easily) retrieved from one or the other. For example, with the EST approach, one easily obtains the probability generating function of customer delay and unfinished work. The SV method does not easily yield these results. On the other hand it is hard to obtain the joint probability generating function of the state description with the EST method. Neither can one easily obtain the total unfinished service time.

Summarising, it is our opinion that either method comes with advantages and disadvantages. However, overall, we slightly favour the EST approach. We will therefore use the EST approach in most of the following sections.

2.3 Two-state Markovian vacations

In this section we extend the results of the preceding sections in the sense that we allow some correlation in the vacation process. That is, a two-state Markovian process is used to model the vacation process. The presented results closely follow our contribution [Fiems et al., 2002b] for both the continue and repeat after interruption operation modes. Results for RAI with resampling have not been published in this context.

Apart from the vacation process, we make the same assumptions as in the preceding sections. Again, we consider a queueing system with infinite storage capacity and a single server which may go on leave. Time is slotted and the numbers of customers

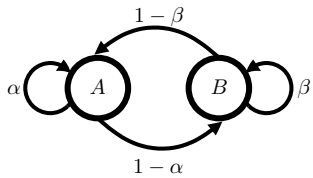


Figure 2.6: Transition diagram of the vacation process.

arriving during the consecutive slots constitute a series of i.i.d. non-negative random variables with common probability generating function $E(z)$. Further, service is synchronised on slot boundaries and the consecutive service times constitute a series of i.i.d. positive random variables with common probability generating function $S(z)$. The characteristics of the vacation process under consideration are described below.

2.3.1 Vacation process

Correlation in the vacation process is modelled by means of a two-state Markov chain. State transition probabilities are depicted in figure 2.6. Whenever the server is in state A, the server is available as opposed to state B where the server is on vacation, in accordance with the definition of A- and B-slots in section 2.1. Alternatively, one may describe this vacation process as an on/off process. The lengths of the consecutive on- or A-periods and off- or B-periods constitute two series of independent and identically geometrically distributed random variables. As before, let $A(z)$ and $B(z)$ denote the common probability generating function of the A- and B-periods respectively, we get,

$$A(z) = \frac{(1-\alpha)z}{1-\alpha z}, \tag{2.87}$$

$$B(z) = \frac{(1-\beta)z}{1-\beta z}. \tag{2.88}$$

Here α and β denote the probabilities that the server remains in the A- or B-state respectively in accordance with the transition probabilities depicted in figure 2.6. It is often convenient to use the fraction of available slots σ and the vacation burstiness factor K instead of the distribution parameters α and β, i.e.,

$$\sigma = \frac{\mu_A}{\mu_A + \mu_B} = \frac{1-\beta}{2-\alpha-\beta}, \tag{2.89}$$

$$K = \sigma\mu_B = (1-\sigma)\mu_A = \frac{1}{2-\alpha-\beta}. \tag{2.90}$$

Given the fraction σ, the parameter K takes values between $\max(\sigma, 1-\sigma)$ and infinity and is a measure for the absolute lengths of the A and B-periods. The vacation burstiness factor K equals 1 for uncorrelated vacations, in which case the server is available with probability $\sigma = \alpha = 1 - \beta$ during each slot. That is, for $K = 1$, the vacation process reduces to the Bernoulli vacation process considered in the preceding sections.

2.3.2 Effective service times

Remember that the effective service time of a customer is defined as the time period – expressed as an integer number of slots – that the system effectively spends on serving this customer, that is, the number of slots between the beginning of the slot where the customer enters the server and the end of the slot where the customer leaves the system. The correlation in the vacation process implies that the consecutive effective service times no longer constitute a series of i.i.d. random variables as a customer's effective service time depends on the state of the server when this customer enters the server. Therefore, it is no longer possible to reduce the model to an equivalent $Geo^X/G/1$ system without vacations by retrieving the probability generating functions of the effective service times. Despite that, the method of the effective service times will still reduce the complexity of the analysis considerably. In particular, we derive expressions for the probability generating functions of the effective service times conditioned on the state of the server during the slot preceding the effective service time. This reduces the analysis of the present system to the analysis of an equivalent queueing system without vacations but with state dependent (effective) service times for all operation modes under consideration.

Continue after interruption

The continue after interruption mode is an operation mode without memory in the sense that during the effective service no record needs to be kept of the complete service time. From a system's point of view, serving the remainder of a customer's service time is equivalent to serving a new customer with service time equal to the remaining service time of the former customer. This implies that the state of the server during the slot preceding the remaining effective service time and the remaining service time completely determine the remaining effective service time distribution.

These remarks lead to a set of recursive equations by conditioning on the state of the server during the first slot of the effective service time. Let $t_A(n|k)$ and $t_B(n|k)$ denote the probability that the (remaining) effective service time of a customer with (remaining) service time equal to k slots equals n slots, given the slot preceding the

start of service is an A- or B-slot respectively, we then get,

$$t_A(n|k) = \alpha t_A(n-1|k-1) + (1-\alpha)t_B(n-1|k) \tag{2.91}$$
$$t_B(n|k) = (1-\beta)t_A(n-1|k-1) + \beta t_B(n-1|k) \tag{2.92}$$

for $n \geq k > 1$. The remaining effective service time is never lower than the remaining customer service time for CAI and therefore $t_A(n|k) = t_B(n|k) = 0$ for $n < k$. Further, a customer leaves the system at the end of the next A-slot if this customer's remaining service time equals one slot. We therefore find:

$$t_A(n|1) = \alpha \qquad \text{for } n = 1, \tag{2.93}$$
$$t_A(n|1) = (1-\alpha)(1-\beta)\beta^{n-2} \qquad \text{for } n > 1. \tag{2.94}$$

Some standard z-transform manipulations then translate the set of recursive equations (2.91)–(2.92) into an equivalent set of the corresponding probability generating functions:

$$T_A(z|k) \triangleq \sum_{n=1}^{\infty} t_A(n|k)z^n = \frac{\alpha + (1-\alpha-\beta)z}{1-\beta z} z T_A(z|k-1), \tag{2.95}$$
$$T_B(z|k) \triangleq \sum_{n=1}^{\infty} t_B(n|k)z^n = \frac{1-\beta}{\alpha + (1-\alpha-\beta)z} T_A(z|k), \tag{2.96}$$

for $k > 1$. Here, $T_A(z|k)$ and $T_B(z|k)$ denote the (conditional) probability generating functions corresponding with $t_A(n|k)$ and $t_B(n|k)$ respectively. Similarly, in accordance with the probabilities $t_A(n|1)$ and $t_B(n|1)$ ($n \geq 1$), we find,

$$T_A(z|1) = \frac{\alpha + (1-\alpha-\beta)z}{1-\beta z} z. \tag{2.97}$$

Equations (2.96), (2.97) and successive application of equation (2.95) then lead to

$$T_A(z|k) = \left(\frac{\alpha + (1-\alpha-\beta)z}{1-\beta z} z \right)^k, \tag{2.98}$$
$$T_B(z|k) = \frac{(1-\beta)z}{1-\beta z} \left(\frac{\alpha + (1-\alpha-\beta)z}{1-\beta z} z \right)^{k-1}. \tag{2.99}$$

Averaging these (conditional) probability generating functions over all possible values of k with respect to the service time distribution then yields expressions for the probability generating functions of the effective customer service times for CAI, given the

state of the server during the slot preceding the start of his effective service:

$$T_A(z) = S\left(\frac{\alpha + (1-\alpha-\beta)z}{1-\beta z}z\right), \tag{2.100}$$

$$T_B(z) = \frac{1-\beta}{\alpha + (1-\alpha-\beta)z} S\left(\frac{\alpha + (1-\alpha-\beta)z}{1-\beta z}z\right). \tag{2.101}$$

For uncorrelated vacations, that is, for $\alpha = \sigma$ and $\beta = 1 - \sigma$, one verifies that the expressions for $T_A(z)$ and $T_B(z)$ both reduce to the expression (2.5) of the effective service times for CAI and Bernoulli vacations as derived in section 2.1. In this particular case, the effective service time distribution does not depend on the state of the server during the slot preceding the effective service time.

For further use, we also retrieve some moments. Using the moment-generating property of generating functions we obtain mean μ_{T_A} and variance ${\sigma_{T_A}}^2$ of the effective service times given that the slot preceding the effective service time is an A-slot as functions of the mean μ_S and variance ${\sigma_S}^2$ of the customer service times and the vacation parameters σ and K:

$$\mu_{T_A} = \frac{\mu_S}{\sigma}, \tag{2.102}$$

$${\sigma_{T_A}}^2 = \frac{{\sigma_S}^2 + \mu_S(2\,K-1)(1-\sigma)}{\sigma^2}. \tag{2.103}$$

In general, the k-th moment of the effective service time (given that the preceding slot is an A-slot) is a function of the vacation parameters K and σ and the moments of the customer service times up to order k. The k-th moment of the effective service time therefore exists for finite K, $0 < \sigma \leq 1$ and when the corresponding moment of the service times exist. Note that the existence of moments of the effective service times in case of CAI does not depend on the vacation parameters σ and K as long as these are positive and finite respectively.

Repeat after interruption

In case of RAI, we have to take into account that service completely restarts after an interruption. We therefore define $t_A(n|k,l)$ as the probability that the remaining effective service time of a customer with total service time equal to k slots equals n slots in case of RAI, given that the remaining service time of the customer equals l slots and given that the slot preceding the remaining effective service time is an A-slot. In case the slot preceding the (remaining) effective service time is a B-slot, one observes that remaining service time and total service time are equal as service is restarted after a B-slot. Therefore, let $t_B(n|k)$ denote the probability that the remaining effective

service time of a customer with service time equal to k slots equals n slots, given that the slot preceding the remaining effective service time is a B-slot.

Conditioning on the state of the server during the first slot of the (remaining) effective service time then yields the following set of recursive equations for these probabilities,

$$t_A(n|k,l) = \alpha t_A(n-1|k,l-1) + (1-\alpha)t_B(n-1|k), \tag{2.104}$$
$$t_B(n|l) = \beta t_B(n-1|l) + (1-\beta)t_A(n-1|l,l-1), \tag{2.105}$$

for $n \geq k \geq l > 1$. The remaining effective service time exceeds the remaining service time and therefore $t_A(n|k,l) = t_B(n|l) = 0$ for $n < l$. Similarly, we obtain the probabilities $t_A(n|k,1)$:

$$t_A(n|k,1) = \alpha \qquad \text{for } n = 1, \tag{2.106}$$
$$t_A(n|k,1) = (1-\alpha)t_B(n-1|k) \qquad \text{for } n > 1. \tag{2.107}$$

Let $T_A(z|k,l)$ and $T_B(z|l)$ denote the probability generating functions corresponding to $t_A(n|k,l)$ and $t_B(n|l)$ respectively, then (2.104) and (2.105) transform into,

$$T_A(z|k,l) \triangleq \sum_{n=1}^{\infty} t_A(n|k,l)z^n = \alpha z T_A(z|k,l-1) + (1-\alpha)\, zT_B(z|k), \tag{2.108}$$
$$T_B(z|l) \triangleq \sum_{n=1}^{\infty} t_B(n|l)z^n = \frac{(1-\beta)z}{1-\beta z} T_A(z|l,l-1), \tag{2.109}$$

for $k \geq l > 1$. Further, equations (2.106) and (2.107) allow us to obtain the probability generating function $T_A(z|k,1)$:

$$T_A(z|k,1) = \alpha z + (1-\alpha)zT_B(z|k). \tag{2.110}$$

The former equation and successive application of equation (2.108) then yields an explicit expression for $T_A(z|k,l)$ in terms of α and $T_B(z|k)$:

$$T_A(z|k,l) = \alpha^l z^l + \frac{(1-\alpha)z}{1-\alpha z}(1-\alpha^l z^l)T_B(z|k). \tag{2.111}$$

Combining the former equation with equation (2.109) then yields explicit expressions

for both $T_A(z|k,l)$ and $T_B(z|k)$ in terms of the vacation parameters α and β,

$$T_A(z|k,l) = \frac{(1-z)(1+(1-\alpha-\beta)z)\alpha^{l+1}z^l + (1-\alpha)(1-\beta)\alpha^k z^{k+1}}{\alpha(1+(1-\alpha-\beta)z)(1-z) + (1-\alpha)(1-\beta)\alpha^k z^{k+1}}, \tag{2.112}$$

$$T_B(z|k) = \frac{\alpha^{k-1}z^k\,(1-\alpha z)\,(1-\beta)}{(1+(1-\alpha-\beta)\,z)\,(1-z) + (1-\alpha)\,(1-\beta)\,\alpha^{k-1}z^{k+1}}. \tag{2.113}$$

The (conditional) probability generating function of the effective service time of a customer with service time equal to k, then equals $T_A(z|k,k)$ or $T_B(z|k)$ depending on the state of the server during the slot preceding effective service. Averaging these (conditional) probability generating functions over all possible values of k with respect to the service time distribution then yields expressions for the probability generating functions of the effective service times for RAI, given the state of the server during the slot preceding the start of effective service:

$$T_A(z) = \sum_{k=1}^{\infty} \frac{s(k)\alpha^{k-1}z^k\,(1-\alpha z)\,(\alpha+(1-\alpha-\beta)\,z)}{(1+(1-\alpha-\beta)\,z)\,(1-z) + (1-\alpha)\,(1-\beta)\,\alpha^{k-1}z^{k+1}}, \tag{2.114}$$

$$T_B(z) = \sum_{k=1}^{\infty} \frac{s(k)\alpha^{k-1}z^k\,(1-\alpha z)\,(1-\beta)}{(1+(1-\alpha-\beta)\,z)\,(1-z) + (1-\alpha)\,(1-\beta)\,\alpha^{k-1}z^{k+1}}. \tag{2.115}$$

One verifies that in the case of uncorrelated vacations ($K = 1$), the former probability generating functions both reduce to the probability generating function (2.14) of a customer's effective service time in case of Bernoulli vacations and RAI. In this particular case, a customer's effective service time does not depend on the state of the server during the slot preceding his effective service time, as expected. Similarly as for the latter system, these expressions are in general not explicit due to the presence of the infinite sum in the right hand side of (2.114). However, one can easily see that the values of all derivatives of these probability generating functions evaluated at $z = 1$ – and as a consequence all moments of the corresponding random variables – can be explicitly calculated in terms of the system parameters. In particular, the mean μ_{T_A} and the variance ${\sigma_{T_A}}^2$ of the effective service time given that the slot preceding the effective service time is an A-slot are given by

$$\mu_{T_A} = \frac{(1-\sigma-K)\left(1-S\left(\frac{1}{\alpha}\right)\right)}{(1-\sigma)\,\sigma} \tag{2.116}$$

and

$$\sigma_{T_A}{}^2 = \frac{\left\{\begin{array}{l} (\sigma+K-1)^2\left(2\,S\left(\frac{1}{\alpha^2}\right)-S\left(\frac{1}{\alpha}\right)^2\right) \\ +2\,KS'\left(\frac{1}{\alpha}\right)(\sigma-1)\,\sigma \\ +S\left(\frac{1}{\alpha}\right)(2\,K-1)\,(\sigma+K-1)\,(\sigma-1)\,\sigma \\ -(\sigma+K-1)\left((1-2\,(1-\sigma)\,\sigma)\,K-(1-\sigma)^2\right) \end{array}\right\}}{(1-\sigma)^2\,\sigma^2}, \tag{2.117}$$

respectively.

In general, the k-th moment of the effective service time (given that the latter is preceded by an A-slot) is a function of the vacation parameters σ and K and of the probability generating function $S(z)$ and its derivatives evaluated at $z = 1/\alpha^i$ $(i = 1 \ldots k)$. One can then observe that for $0 < \sigma \leq 1$ and finite K the k-th moment exists for $\alpha > 1/\sqrt[k]{R_S}$ and does not exist for $\alpha < 1/\sqrt[k]{R_S}$. For $\alpha = 1/\sqrt[k]{R_S}$, the existence depends on the behaviour of $S(z)$ and its derivatives for $z = R_S$. Here, R_S denotes the radius of convergence of the probability generating function $S(z)$ as before. Note, that α is function of both σ and K. Therefore, as opposed to CAI, requiring the existence of the moments of the effective service times constrains both σ and K.

Repeat after interruption with resampling

As service time is resampled after an interruption, RAI,wr is also an operation mode without memory in the sense that no record needs to be kept of the complete customer service time during this customer's effective service time. There is again from a system's point of view, no difference between serving the remainder of a customer's service time or serving a new customer with service time equal to the remaining service time of the former customer. Let $t_A(n|k)$ denote the probability that a customer's remaining effective service time takes n slots given that the slot preceding the effective service time is an A-slot and given that this customer would complete service in k slots under the assumption that there are no vacations. Similarly, let $t_B(n)$ denote the probability that a customer's effective service time equals n slots given that the slot preceding the effective service time is a B-slot. Conditioning on the state of the server during the first slot of the effective service time then yields,

$$t_A(n|k) = \alpha t_A(n-1|k-1) + (1-\alpha)t_B(n-1), \tag{2.118}$$

$$t_B(n) = \beta t_B(n-1) + (1-\beta)\sum_{j=2}^{\infty} s(j)t_A(n-1|j-1), \tag{2.119}$$

for $n, k > 1$. Further, a customer leaves the system at the end of a slot if his remaining service time equals one at the beginning of this slot and if the server is available during this slot. Therefore we get: $t_A(1|1) = \alpha$, $t_A(1|k) = 0$ for $k > 1$ and $t_B(1) = (1-\beta)s(1)$. The server may leave for a vacation when a customer's remaining service time equals one slot: $t_A(n|1) = (1-\alpha)t_B(n-1)$.

Using some standard z-transform manipulations, we obtain following set of equations for the probability generating functions $T_A(z|k)$ and $T_B(z)$ corresponding to the probabilities $t_A(n|k)$ and $t_b(n)$ respectively:

$$T_A(z|k) = \alpha z T_A(z|k-1) + (1-\alpha) z T_B(z), \tag{2.120}$$

$$T_B(z) = \beta z T_B(z) + (1-\beta) z s(1) + (1-\beta) z \sum_{j=2}^{\infty} s(j) T_A(z|j-1), \tag{2.121}$$

for $k > 1$ and $T_A(z|1) = \alpha z + (1-\alpha) T_B(z)$. The former expression and successive application of equation (2.120) then yields

$$T_A(z|k) = \frac{(\alpha z)^k (1 - \alpha z - T_B(z) z (1-\alpha)) + T_B(z) z (1-\alpha)}{1 - \alpha\, z}. \tag{2.122}$$

We can now substitute of the former expression into equation (2.121) and solve for $T_B(z)$. We find following expression for $T_B(z)$:

$$T_B(z) = \frac{S(\alpha z)(1-\alpha z)(1-\beta)}{\alpha\,(1 + (1-\alpha-\beta)\,z)\,(1-z) + (1-\alpha)\,(1-\beta)\,z S(\alpha z)}. \tag{2.123}$$

On the other hand, substitution of the former result into equation (2.122) and averaging over all possible customer service times with respect to their probabilities yields

$$T_A(z) = \frac{S(\alpha z)\,(1-\alpha z)\,(\alpha + (1-\alpha-\beta)\,z)}{\alpha\,(1 + (1-\alpha-\beta)\,z)\,(1-z) + (1-\alpha)\,(1-\beta)\,z S(\alpha z)}. \tag{2.124}$$

Using the moment-generating property of generating functions, we may again derive various moments. In particular, the mean μ_{T_A} and the variance ${\sigma_{T_A}}^2$ of the effective service times given that the slot preceding the effective service time is an A-slot are given by,

$$\mu_{T_A} = \frac{(1-\sigma-K)\,(S(\alpha)-1)}{S(\alpha)\,(1-\sigma)\,\sigma}, \tag{2.125}$$

and

$$\sigma_{T_A}{}^2 = \frac{\left\{\begin{array}{l}(K+\sigma-1)^2\,(2\,S'(\alpha)\sigma\,(\sigma-1)+K)\\ +\,(K+\sigma-1)\left((1-\sigma)^2-(1-2\,\sigma(1-\sigma))K\right)KS(\alpha)^2\\ -\,(K+\sigma-1)\,(2\,K-1)\,\sigma\,(1-\sigma)\,KS(\alpha)\end{array}\right\}}{\sigma^2\,(1-\sigma)^2\,KS(\alpha)^2}, \quad (2.126)$$

respectively. In accordance with our findings for the Bernoulli model, for $0 < \sigma < 1$ and finite K, all moments of the effective service times exist whereas for $\sigma = 1$ and finite K existence of the n-th moment requires existence of the n-th moment of the service time. That is, requirements are more stringent in the absence of vacations.

2.3.3 Queue content

We now use the former results to retrieve the probability generating function of the queue content at departure epochs and at random slot boundaries. We first consider the queue content at customer departure times.

Queue content at departure epochs

As before, let $V_d^{(k)}$ denote the queue content at the k-th departure epoch. That is, at the beginning of the slot following the departure slot of the k-th customer. For positive $V_d^{(k)}$, service of the $(k+1)$-th customer can start immediately. This implies that – as the previous slot was an A slot since there was a departure – it will take T_A slots to the next departure, with T_A a random variable representing the effective service time of a customer given his service was preceded by an A-slot, and whose probability generating function is given by (2.100), (2.114) or (2.124) depending on the operation mode under consideration. The queue content at the $(k+1)$-th departure epoch therefore relates to the content at the k-th departure epoch as follows:

$$V_d^{(k+1)} = V_d^{(k)} - 1 + \sum_{j=1}^{T_A} E^{(j)}, \quad (2.127)$$

for $V_d^{(k)} > 0$ and where $E^{(j)}$ denotes the number of customers arriving in the system during the j-th slot of the effective service time of the $(k+1)$-th customer.

On the other hand, if the queue is empty after the departure of the k-th customer, the next customer will not be served immediately as the latter still has to arrive in the system. Consider therefore the first slot where there is at least one customer arrival following the departure slot of the k-th customer. Let $\tilde{E}^{(k)}$ denote the number of customer arrivals during this slot and let $\tilde{Q}^{(k)}$ denote the state of the server during this

slot (A or B). As service of the $(k+1)$-th customer starts in the slot following this slot and as the effective service time of this customer is described by the random variable $T_{\tilde{Q}^{(k)}}$, the queue content at the $(k+1)$-th departure epoch is given by:

$$V_d^{(k+1)} = \tilde{E}^{(k)} - 1 + \sum_{j=1}^{T_{\tilde{Q}^{(k)}}} E^{(j)}, \tag{2.128}$$

for $V_d^{(k)} = 0$. The independence in the arrival process and the fact that the server is available during a customer's departure slot assure that $\tilde{E}^{(k)}$ and $T_{\tilde{Q}^{(k)}}$ are independent random variables with distributions independent of k. Let $\tilde{E}(z)$ and $T_{\tilde{Q}}(z)$ denote the corresponding probability generating functions. From the system equations (2.127) and (2.128), we then obtain the following relation between the probability generating functions $V_d^{(k)}(z)$ and $V_d^{(k+1)}(z)$ of the queue content at the k-th and $(k+1)$-th departure epochs,

$$\begin{aligned} V_d^{(k+1)}(z) = &\left(V_d^{(k)}(z) - V_d^{(k)}(0)\right) T_A(E(z))\frac{1}{z} \\ &+ V_d^{(k)}(0)\tilde{E}(z)T_{\tilde{Q}}(E(z))\frac{1}{z}. \end{aligned} \tag{2.129}$$

Under the assumption that the system under consideration reaches steady state (weak stability), let $V_d(z) = \lim_{k\to\infty} V_d^{(k)}(z)$ denote the steady-state probability generating function of the queue content at departure epochs. One can show – we will prove this assertion in a more general context in section 2.4 – that the system reaches steady state whenever the effective load $\tilde{\rho}$ is less than one,

$$\tilde{\rho} = \mu_E \mu_{T_A} < 1. \tag{2.130}$$

Recall that expressions for the mean effective service times μ_{T_A} are displayed in equations (2.102), (2.116) and (2.125) for the CAI, RAI and RAI,wr operation modes respectively. Equation (2.129) then easily yields the following expression for the probability generating function $V_d(z)$ of the queue content at departure epochs in steady state,

$$V_d(z) = V_d(0)\,\frac{\tilde{E}(z)\,T_{\tilde{Q}}(E(z)) - T_A(E(z))}{z - T_A(E(z))}. \tag{2.131}$$

The unknown constant $V_d(0)$ then follows from the normalisation condition $V_d(1) = 1$,

$$V_d(0) = \frac{1 - \mu_{T_A}\,\mu_E}{\mu_{\tilde{E}} + \mu_{T_{\tilde{Q}}}\,\mu_E - \mu_{T_A}\,\mu_E}. \tag{2.132}$$

Here, $\mu_{\tilde{E}}$ and $\mu_{T_{\tilde{Q}}}$ denote the mean values of $\tilde{E}$ and $T_{\tilde{Q}}$ respectively. We now derive expressions for the corresponding probability generating functions.

First consider the probability generating function $\tilde{E}(z)$. From a stochastic point of view, the only difference between a random slot and the first slot with arrivals following a departure slot is that in the latter there is at least one arrival. The corresponding generating function is given in equation (1.11), that is,

$$\tilde{E}(z) = \frac{E(z) - E(0)}{1 - E(0)}. \tag{2.133}$$

Finally, the probability generating function $T_{\tilde{Q}}(z)$ can be retrieved as follows. Define $\theta(n)$ as the probability that the n-th slot after an A-slot is an A-slot. Explicit expressions for these values are easily derived by means of the following recursive equation,

$$\theta(n) = \alpha\,\theta(n-1) + (1-\beta)\,(1-\theta(n-1)) \tag{2.134}$$

for $n > 0$, with boundary condition $\theta(0) = 1$. The former expression then leads to an expression for the corresponding z-transform $\Theta(z)$,

$$\Theta(z) \triangleq \sum_{n=1}^{\infty} \theta(n)\, z^n = \frac{\alpha + z\,(1-\alpha-\beta)}{1 + z\,(1-\alpha-\beta)}\,\frac{z}{1-z}. \tag{2.135}$$

Note that $\Theta(z)$ is a z-transform of a series and not a probability generating function.

Due to the nature of the arrival process, the first customer arrival after some (tagged) slot occurs during the n-th slot after this tagged slot with probability

$$\kappa(n) = E(0)^{n-1}\,(1 - E(0)), \tag{2.136}$$

for $n > 0$. In particular, the first slot with arrivals following the departure slot of a customer which leaves behind an empty system, is the n-th slot following this departure slot with probability $\kappa(n)$. Furthermore, this slot is an A-slot with probability $\theta(n)$ as the server is available during the last slot of the effective service time of customers. Summation over all possible values of n yields an expression for the probability ϕ that the server is available during the first slot where a customer arrival occurs after the departure of some customer,

$$\begin{aligned} \phi &\triangleq \sum_{n=1}^{\infty} \theta(n)\,\kappa(n) = \Theta(E(0))\,\frac{1-E(0)}{E(0)} \\ &= \frac{\alpha + E(0)\,(1-\alpha-\beta)}{1 + E(0)\,(1-\alpha-\beta)}. \end{aligned} \tag{2.137}$$

The effective service time of the first customer arriving during this slot is therefore preceded by an A-slot (B-slot) with probability ϕ (with probability $1-\phi$). The corresponding probability generating function $T_{\tilde{Q}}(z)$ of this customer's effective service time is therefore given by

$$T_{\tilde{Q}}(z) = \phi\, T_A(z) + (1-\phi)\, T_B(z). \tag{2.138}$$

From equations (2.100), (2.101), (2.114), (2.115), (2.123) and (2.124), one observes that $T_A(z)$ and $T_B(z)$ are similarly related for all operation modes under consideration,

$$T_B(z) = \frac{1-\beta}{\alpha + (1-\alpha-\beta)\, z}\, T_A(z). \tag{2.139}$$

The latter relation can also be retrieved as follows. Consider the decomposition of the effective service time in two components: the number of slots until the first available slot and the remaining effective service time. Both components are independent random variables. It is clear that the first component (and its pgf) does not depend on the operation mode whereas the second component does not depend on the state of the server during the slot preceding the effective service as the last slot of the first component is by definition an A-slot. Let $\Upsilon_A(z)$ and $\Upsilon_B(z)$ denote the probability generating functions of the first component given the state during the slot preceding the effective service and let $\Lambda_{(mode)}(z)$ denote the pgf of the second component only depending on the operation mode, then

$$T_A(z) = \Upsilon_A(z)\, \Lambda_{(mode)}(z), \tag{2.140}$$
$$T_B(z) = \Upsilon_B(z)\, \Lambda_{(mode)}(z), \tag{2.141}$$

with

$$\Upsilon_A(z) = \alpha\, z + (1-\alpha)\, z\, \frac{(1-\beta)z}{1-\beta z}, \tag{2.142}$$
$$\Upsilon_B(z) = \frac{(1-\beta)z}{1-\beta z}. \tag{2.143}$$

The former expression – an A-slot precedes the effective service time – follows from the fact that the first A-slot is either the first slot of the effective service time (with probability α) or the first slot after a geometrically distributed vacation (with probability $(1-\alpha)$). The latter expression – a B-slot precedes the effective service time – follows from the fact that the first A-slot follows when the vacation ends. Elimination of $\Lambda_{(mode)}(z)$ from the equations above then yields expression (2.139).

Plugging this expression and equation (2.137) into equation (2.138) then yields,

$$T_{\tilde{Q}}(z) = \frac{\left\{ \begin{array}{l} (1-\alpha-\beta)((\alpha + z\,(1-\alpha-\beta))E(0) + z\alpha) \\ +1-(1-\alpha)(\alpha+\beta) \end{array} \right\}}{(1+E(0)(1-\alpha-\beta))(\alpha+z(1-\alpha-\beta))}\, T_A(z). \tag{2.144}$$

Bringing everything together, we get the following expression for the probability generating function $V_d(z)$ of the queue content at departure epochs in steady state,

$$V_d(z) = \sigma\, \frac{1-\tilde{\rho}}{\mu_E}\, \frac{1+E(z)(1-\alpha-\beta)}{\alpha+E(z)(1-\alpha-\beta)}\, \frac{T_A(E(z))(1-E(z))}{T_A(E(z))-z}. \tag{2.145}$$

Queue content at random slot boundaries

The independence in the arrival process under consideration implies that the probability generating functions of the queue content at departure epochs $V_d(z)$ and at random slot boundaries $V_r(z)$ are related as

$$V_r(z) = \frac{V_d(z)(z-1)\mu_E}{E(z)-1}. \tag{2.146}$$

The former relation was derived in section 1.2. The probability generating function of queue content at random slot boundaries is then given by

$$V_r(z) = \sigma\,(1-\tilde{\rho})\, \frac{1+E(z)(1-\alpha-\beta)}{\alpha+E(z)(1-\alpha-\beta)}\, \frac{T_A(E(z))(1-z)}{T_A(E(z))-z}. \tag{2.147}$$

Moments

Finally, taking the appropriate derivatives of (2.145) and (2.147) yields expressions for the various moments of the queue content at departure epochs and at random slot boundaries respectively. In particular, the mean queue content at departure epochs μ_{V_d} and at random slot boundaries μ_{V_r} are given by

$$\begin{aligned} \mu_{V_d} = & \frac{{\sigma_E}^2 + {\mu_E}^3\, {\sigma_{T_A}}^2 - \mu_E\,(1-\tilde{\rho})\,(1-\tilde{\rho}-\mu_E)}{2\,(1-\tilde{\rho})\,\mu_E} \\ & + (K-1)\frac{(1-\sigma)\,\mu_E}{\sigma} \end{aligned} \tag{2.148}$$

and

$$\mu_{V_r} = \frac{\tilde{\rho}(1-\tilde{\rho}) + {\mu_E}^2 {\sigma_{T_A}}^2 + \mu_{T_A} {\sigma_E}^2}{2\,(1-\tilde{\rho})} + (K-1)\frac{(1-\sigma)\,\mu_E}{\sigma} \tag{2.149}$$

respectively. Expressions for the mean and the variance of the effective service times for the different operation modes can be found in the preceding section. For finite K and positive σ, the moments exist – the system is then strongly stable – whenever the system is weakly stable and the means and the variances of the effective service times and the number of customer arrivals in a slot exist. The existence of the variance of the effective service times was discussed in the preceding section for the different operation modes under consideration. We will numerically investigate conditions for strong stability in section 2.3.5.

2.3.4 Unfinished Work and Customer Delay

Unfinished work

As before, the unfinished work at a given time instant is defined as the number of slots it would take to empty the queue under the assumption that there are no new arrivals. The latter again includes the B-slots during which the server is on vacation, and (in case of RAI and RAI,wr) slots lost due to uncompleted interrupted service trials. Note that according to this definition, the unfinished work equals zero if and only if the system is empty. Consider now a random slot k, and define $U^{(k)}$ as the unfinished work at the beginning of this slot. If the queue is not empty, the unfinished work is diminished by 1 and for each customer arriving in this slot, an additional number of slots equal to the effective service time of this customer is required to empty the queue, i.e., the unfinished work at the beginning of slot $(k+1)$ relates to the unfinished work at the beginning of slot k as follows:

$$U^{(k+1)} = U^{(k)} - 1 + \sum_{j=1}^{E} T_A^{(j)}, \tag{2.150}$$

for $U^{(k)} > 0$. Here, E denotes the number of arrivals in slot k and $T_A^{(j)}$ the effective service time of the j-th customer arriving in slot k. The latter series of random variables share the common probability generating function $T_A(z)$. This follows from the fact that each of these customers start service after the departure slot of the preceding customer. By definition, a customer receives service during his departure slot. That is, a departure slot is an A-slot.

In the case that the system is empty at the beginning of slot k, it remains empty if there are no arrivals. If there is at least one arrival during slot k, the unfinished work at the

beginning of slot $(k+1)$ equals the sum of the effective service time of this customer and of the effective service times of other customers arriving during slot k,

$$U^{(k+1)} = T_{\hat{Q}^{(k)}} + \sum_{j=2}^{E} T_A^{(j)}, \tag{2.151}$$

for $U^{(k)} = 0$ and $E > 0$. Here $T_{\hat{Q}^{(k)}}$ denotes the effective service time of the first customer arriving during slot k ($\hat{Q}^{(k)}$ denotes the state of the server during slot k) , E denotes the number of arrivals during slot k and $T_A^{(j)}$ denotes the effective service time of the j-th customer arriving during slot k. The latter series again share the probability generating function $T_A(z)$ as all but the first arriving customer start service after the departure slot (A-slot) of the preceding customer.

The system equations (2.150) and (2.151) and some standard z-transform manipulations then yield the following relation between the probability generating functions $U^{(k)}(z)$ and $U^{(k+1)}(z)$ of the unfinished work at the k-th and $(k+1)$-th slot boundary,

$$\begin{aligned} U^{(k+1)}(z) = {} & (U^{(k)}(z) - U^{(k)}(0))E(T_A(z))\frac{1}{z} + U^{(k)}(0)E(0) \\ & + U^{(k)}(0)\frac{T_{\hat{Q}^{(k)}}(z)}{T_A(z)}(E(T_A(z)) - E(0)). \end{aligned} \tag{2.152}$$

Now assume again that the system under consideration reaches steady state and let $U(z) = \lim_{k\to\infty} U^{(k)}(z)$ denote the probability generating function of the unfinished work in steady state. From the former equation, we then easily find the following expression for $U(z)$,

$$U(z) = U(0)\frac{E(T_A(z))(zT_{\hat{Q}}(z) - T_A(z)) + E(0)z(T_A(z) - T_{\hat{Q}}(z))}{T_A(z)(z - E(T_A(z)))}. \tag{2.153}$$

Here $T_{\hat{Q}}(z)$ denotes the probability generating function of the effective service time of a customer that arrives in an empty system. As the periods that the system remains empty (empty periods) are geometrically distributed, one easily shows that a random slot during which the system is empty, is the n-th slot of an empty period with probability $\kappa(n)$ as given by (2.136). In accordance with equation (2.137), we then find that the probability that the server is available during the arrival slot of a customer that arrives in an empty system equals ϕ. Therefore, $T_{\hat{Q}}(z)$ equals $T_{\tilde{Q}}(z)$ as displayed in equations (2.138) and (2.144).

Using equation (2.144) and the fact that an empty queue implies zero unfinished work

and vice versa, i.e. $U(0) = V_r(0)$, equation (2.153) transforms into,

$$U(z) = \frac{\sigma\,(1-\tilde{\rho})(1-z)}{\alpha + E(0)(1-\alpha-\beta)} \times \frac{\left\{\begin{array}{l}(1-\alpha-\beta)(\alpha+z(1-\alpha-\beta))E(0)E(T_A(z)) \\ +\,\alpha\,(1+z(1-\alpha-\beta))E(T_A(z)) + E(0)z(1-\alpha)(1-\alpha-\beta)\end{array}\right\}}{(\alpha+z(1-\alpha-\beta))(E(T_A(z))-z)}. \quad (2.154)$$

Customer delay

Consider now a particular (tagged) customer. The delay D – the number of slots between the end of his arrival slot and the end of his departure slot – of this customer is the sum of the unfinished work U at the beginning of this customer's arrival slot diminished by one slot if there is such unfinished work, of the effective service times of all customers that arrive in the same slot but before the tagged customer and of the tagged customer's effective service time. In case the system is not empty at the beginning of the tagged customer's arrival slot, we get,

$$D = U - 1 + \sum_{j=1}^{\check{E}+1} T_A^{(j)}. \quad (2.155)$$

Here $T_A^{(j)}$ denotes the effective service time of the j-th customer arriving in the tagged customer's arrival slot whereas $\check{E}$ denotes the number of customers that arrive in the same slot as the tagged customer and that are served before the tagged customer. The former series share the common probability generating function $T_A(z)$ as they immediately start service after the departure of the preceding customer. The probability generating function corresponding to the latter is given by,

$$\check{E}(z) = \frac{E(z)-1}{(z-1)\,\mu_E}, \quad (2.156)$$

as shown in section 1.2. In case the system is empty at the beginning of the tagged customer's arrival slot, we get,

$$D = T_{\hat{Q}} + \sum_{j=1}^{\check{E}} T_A^{(j+1)}, \quad (2.157)$$

where $T_{\hat{Q}}$ denotes the effective service time of a customer that arrives in an empty system. The corresponding probability generating function $T_{\hat{Q}}(z)$ was derived in the preceding subsection.

Due to the independence in the arrival process, the unfinished work at the beginning of a tagged customer's arrival slot and the unfinished work at the beginning of a random slot boundary share the same probability generating function $U(z)$. Equations (2.155) and (2.157) and some standard z-transform manipulations then yield

$$D(z) = (U(z) - U(0))\,\check{E}(T_A(z))\,\frac{T_A(z)}{z} + U(0)\check{E}(T_A(z))\,T_{\hat{Q}}(z), \tag{2.158}$$

which further simplifies to,

$$D(z) = \sigma\,\frac{1-\tilde{\rho}}{\mu_E}\,\frac{1+z\,(1-\alpha-\beta)}{\alpha+z\,(1-\alpha-\beta)}\,\frac{(1-z)\,T_A(z)}{T_A(z)-1}\,\frac{1-E(T_A(z))}{z-E(T_A(z))}. \tag{2.159}$$

Moments

The various moments of unfinished work and customer delay can then be determined by taking the appropriate derivatives of (2.154) and (2.159) respectively. In particular, mean unfinished work and mean customer delay are given by,

$$\begin{aligned}\mu_U = &\frac{{\mu_T}^2{\sigma_E}^2 + \tilde{\rho}\,(1-\tilde{\rho}) + \mu_E{\sigma_T}^2}{2\,(1-\tilde{\rho})} \\ &+ \frac{(K-1)\,(\phi - (\tilde{\rho}\,\phi - \tilde{\rho} + 1)\,\sigma)}{\phi\,\sigma} - \frac{(1-\tilde{\rho})\,(\sigma-\phi)}{\phi}\end{aligned} \tag{2.160}$$

and

$$\mu_D = \frac{\tilde{\rho}(1-\tilde{\rho}) + {\mu_E}^2{\sigma_{T_A}}^2 + \mu_{T_A}{\sigma_E}^2}{2\,\mu_E\,(1-\tilde{\rho})} + (K-1)\frac{(1-\sigma)}{\sigma} \tag{2.161}$$

respectively. Here, we also used equation (2.137) to simplify the results. For an uncorrelated vacation process, we get $K = 1$ and $\phi = \sigma$. Therefore, only the first term of the former expressions remains, in accordance with our earlier results for the system with Bernoulli vacations. Note that mean customer delay may also be retrieved using the discretised equivalent of Little's result (see section 1.2.5).

2.3.5 Numerical example

We here investigate the conditions for strong stability by means of some numerical examples. More extensive numerical examples are again delayed until section 2.5 where we use vacation queueing systems to assess performance of preemptive priority queueing systems.

Recall that the vacation system is strongly stable if all of the following conditions are satisfied:

- the steady-state condition (2.130) is satisfied.
- the mean and the variance of the effective service times for the operation mode under consideration exist.
- the mean and the variance of the number of customer arrivals in a slot exist.
- the server is available from time to time, i.e., $0 < \sigma \leq 1$ and the vacation burstiness factor is finite, that is, $\max(\sigma, 1-\sigma) \leq K < \infty$. Recall that the lower bound for the burstiness factor K follows from the definition of this factor.

We may now investigate for which (σ, K) pairs the system is strongly stable. We here assume that the mean and the variance of the number of customer arrivals in a slot as well as the mean and the variance of the customer service times exist.

For the continue after interruption mode, the former assumptions imply that for all $\max(\sigma, 1-\sigma) \leq K < \infty$ and $0 < \sigma \leq 1$, mean and variance of the effective service times exist. Substitution of the expression for the mean effective service time for CAI (2.102) in the steady-state condition (2.130) further yields a lower bound for σ independent of the value of K: $\sigma > \mu_E\, \mu_S$.

For the repeat after interruption mode, the variance of the effective service time exists for $\alpha > 1/\sqrt{R_S}$ where R_S denotes the radius of convergence of the probability generating function $S(z)$. Solving equations (2.89) and (2.90) for α and β, one easily observes that for fixed σ, α increases for increasing values of K and converges to 1 for K tending to infinity. As such, for all $0 < \sigma \leq 1$ there exists a lower bound for K such that $\alpha > 1/\sqrt{R_S}$,

$$\max\left(\sigma, \frac{1-\sigma}{1-1/\sqrt{R_S}}\right) < K < \infty. \tag{2.162}$$

Here, we also take into account that K is by definition bounded by $\max(\sigma, 1-\sigma)$. The expression (2.116) for the mean effective service time for RAI further reveals that for any fixed σ, the mean effective service time exists for $R_S(1-\sigma)/(R_S-1) < K < \infty$, decreases for increasing values of K and converges to the mean effective service time for CAI (for the same customer service time characteristics) for K tending to infinity. This implies that for any $\mu_E\, \mu_S < \sigma \leq 1$, there exists a lower bound K_0 for K such that the steady-state condition (2.130) is satisfied (the system is weakly stable) for all $K > K_0$. The latter bound depends on σ, on μ_E and on the probability generating function of the customer service time $S(z)$. Combining this observation with the inequality (2.162), one observes that for any $\sigma \in (\mu_E\, \mu_S, 1]$, there exists a lower bound for K such that the system is strongly stable in the case of the RAI operation mode.

Finally, for the repeat after interruption with resampling mode, given the assumptions, the variance of the effective service times always exist for all $0 < \sigma \leq 1$ and $\max(\sigma, 1-\sigma) \leq K < \infty$. As for RAI, for some fixed server availability σ the mean effective service time decreases and converges to the mean effective service time for CAI for increasing values of the burstiness factor K. Again, this implies that for all $\mu_E\,\mu_S < \sigma \leq 1$, there exist some lower bound K_1 such that the steady-state condition (2.130) is satisfied (weak stability) for all $K > K_1$. Again, this lower bound depends on σ, on μ_E and on the probability generating function of the customer service time $S(z)$.

Figures 2.7, 2.8 and 2.9 all depict the lower bound for the burstiness factor K versus the server availability σ such that the system is strongly stable. That is, these figures depict the boundaries of the region in the (σ, K) plane where the system is strongly stable. All figures assume that the mean number of arrivals in a slot equals $\mu_E = 0.05$. The customer service times follow a shifted geometrical distribution in figure 2.7, a deterministic distribution in figure 2.8 and a shifted symmetrical binomial distribution in figure 2.9. The reader is referred to appendix A for the characteristics of these distributions. The distributions under consideration are completely characterised by their mean value. Different mean values are considered as depicted.

In the case of the shifted geometrical distribution (figure 2.7), CAI and RAI,wr operate equivalently due to the lack of memory of the geometrical distribution. Therefore, these modes share the same stability region. One may also observe that for large K, the constraints on σ for RAI converge to those for CAI and RAI,wr. This follows from the fact that the available periods are long for large K which implies that service repetitions due to interruptions are relatively scarce.

For deterministic service times (figure 2.8), RAI and RAI,wr operate equivalently and therefore share the same stability region. We here note that for large K, the constraints on σ for RAI and RAI,wr converge to those for CAI which again follows from the fact that service repetitions caused by interruptions are scarce for large values of K. Further, the stability region for CAI does not depend on the complete distribution of the customer service times but only on the mean. Therefore, the same stability regions are obtained here as in figure 2.7. For RAI, existence of the variance of the effective service time is always assured as the radius of convergence of the probability generating function $S(z)$ is infinite for deterministic service times.

Finally, for shifted symmetrically binomial service times (figure 2.9), all modes operate differently. The stability region for CAI is larger than the one for RAI,wr, which is in turn larger than the one for RAI. Again, constraints on σ for large K are almost equal for all modes which again follows from the fact that service interruptions (and repetitions) are scarce for large K. As for deterministic service times, the radius of convergence of the probability generating function of the customer service times is infinite. Therefore, the variance of the effective service times always exists.

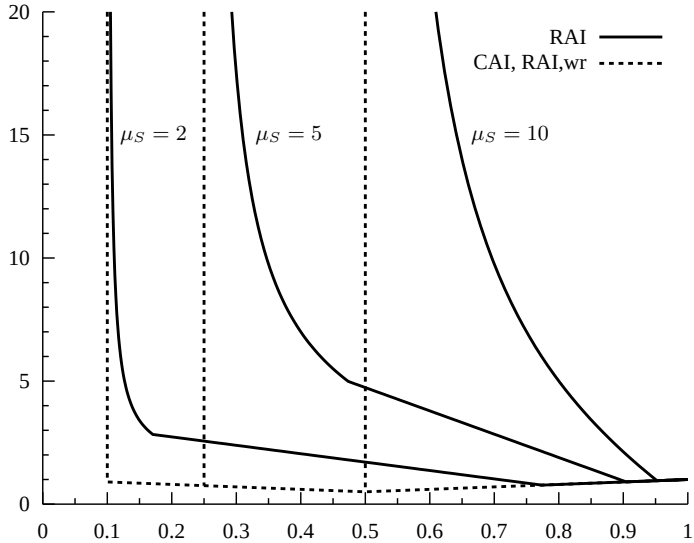


Figure 2.7: The minimal burstiness factor K for strong stability vs. the server availability σ for different values of the mean service time. (shifted geometrically distributed service times, the mean number of arrivals during a slot: $\mu_E = 0.05$.)

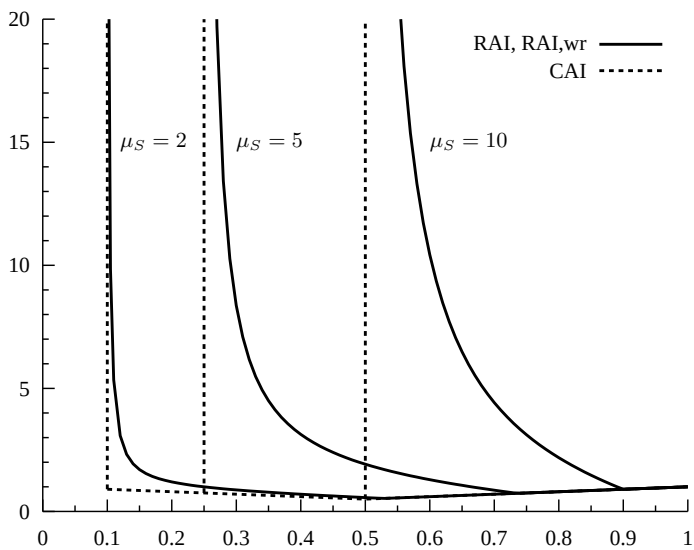


Figure 2.8: The minimal burstiness factor K for strong stability vs. the server availability σ for different values of the mean service time. (deterministic service times, the mean number of arrivals during a slot: $\mu_E = 0.05$.)

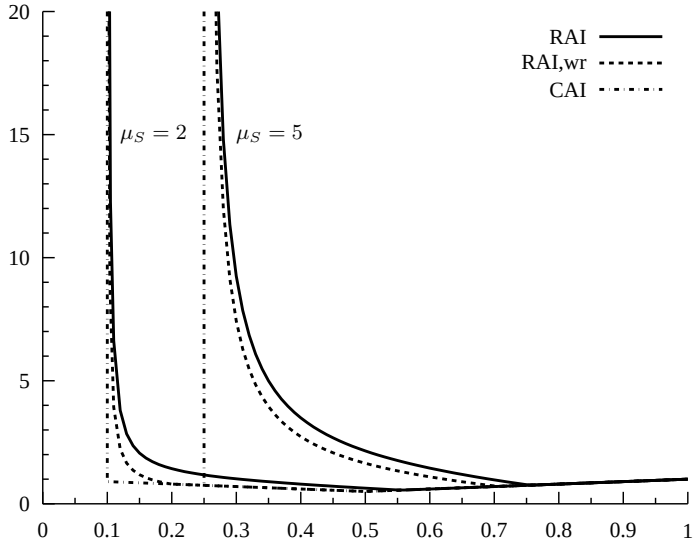


Figure 2.9: The minimal burstiness factor K for strong stability vs. the server availability σ for different values of the mean service time. (shifted symmetric binomial service times, the mean number of arrivals during a slot: $\mu_E = 0.05$.)

2.4 Generally distributed vacations

In the previous section, available periods and vacations obey a geometric distribution. In the present section, we remove the restriction that the vacations are geometrically distributed. Vacations here follow some general distribution. We still assume that the consecutive available periods share a common geometric distribution. This implies that a vacation starts after any available slot with a fixed probability.

Again we consider the continue after interruption mode, the repeat after interruption mode and the repeat after interruption mode with resampling. The results for CAI and RAI are based on our contribution [Fiems et al., 2003], whereas results for RAI,wr are not published yet in this context. The analysis presented in [Fiems et al., 2003] is comparable with the approach followed in the preceding section. Here we follow a different approach. We combine known results for the exhaustive multiple vacation queueing system (see further) and effective service times. This approach will more easily yield stability conditions as we will see further.

Apart from the vacation process, we make the same assumptions as in the preceding sections. That is, we consider a queueing system with infinite storage capacity and a single server which may go on leave from time to time. Time is slotted and the numbers of customers arriving during the consecutive slots constitute a series of i.i.d. non-negative random variables with common probability generating function $E(z)$.

Service is synchronised on slot boundaries and the consecutive service times constitute a series of i.i.d. positive random variables with common probability generating function $S(z)$. The characteristics of the vacation process under consideration are described below.

2.4.1 Vacation process

The server leaves for a vacation from time to time. The vacation process is modelled by two series of i.i.d. positive random variables denoting the lengths of the consecutive available periods and vacation periods. Let $A(z)$ and $B(z)$ denote the common probability generating functions of the series corresponding to the consecutive available and vacation periods respectively. Further, let $b(n)$ ($n > 0$) denote the common probability mass function of the vacation periods. We make no assumptions regarding the common distribution of the vacation periods but assume that the consecutive available periods share a common shifted geometrical distribution. The corresponding probability generating function $A(z)$ is given by

$$A(z) = \frac{(1-\alpha)\,z}{1-\alpha\,z}, \tag{2.163}$$

with α the probability that an A-period continues after an A-slot. As before, let σ denote the fraction of slots that the server is available. One easily verifies that the latter is given by

$$\sigma = \frac{1}{1+(1-\alpha)\,\mu_B}. \tag{2.164}$$

One observes that $\alpha = 1$ implies $\sigma = 1$ and also vice versa as the mean vacation length is by definition non-zero. In this case, the server is always available.

2.4.2 Effective service times

We first concentrate on expressions for the probability generating functions of the effective service times. Recall that the effective service time of a customer is defined as the number of slots between the beginning of the slot where the customer enters the server and the end of the slot where the customer leaves the system. Due to the nature of the vacation process and the service times, the effective service times of consecutive customers constitute a series of independent positive random variables, with distributions depending on the state of the server – described by the availability of the server (A or B) together with the number of remaining B-slots in case the server is unavailable – during the slot preceding the start of the effective service time of the customer and on the operation mode under consideration. We here focus in particular

on the probability generating functions of the effective service times given that the slot preceding the effective service times is an A-slot for the different operation modes under consideration.

Continue after interruption

Let $t(n|k)$ denote the probability that the effective service time of a customer with a k-slot service time equals n slots given that the slot preceding the effective service time is an A-slot. As in section 2.3, we use the fact that the continue after interruption mode is a memoryless operation mode. From the point of view of the system, once a customer has received service during a slot, there is no difference between completing this customer's remaining service and serving a new customer with a service time which is one slot smaller. Therefore, conditioning on whether or not the server leaves for a vacation during the first slot of the effective service time and on the length of this vacation if the server leaves for a vacation yields

$$t(n|k) = \alpha\, t(n-1|k-1) + (1-\alpha) \sum_{j=1}^{n-k} b(j) t(n-j-1|k-1) \tag{2.165}$$

for $n \geq k > 1$. As a customer's effective service time is at least as long as his service time for the CAI operation mode, we also find that the probability $t(n|k)$ equals 0 for $n < k$. Consider now the (remaining) effective service time given that the slot preceding the (remaining) effective service is an A-slot of a customer who's (remaining) service time equals 1 slot. The latter equals 1 slot if the first (and only) effective service time slot is an A-slot and equals the length of a vacation augmented with a single slot if this is not the case. We therefore find:

$$t(n|1) = \alpha \qquad \text{for } n = 1, \tag{2.166}$$

$$t(n|1) = (1-\alpha)\, b(n-1) \qquad \text{for } n > 1. \tag{2.167}$$

Let $T(z|k)$ denote the conditional probability generating function corresponding to $t(n|k)$, that is,

$$T(z|k) = \sum_{n=k}^{\infty} t(n|k)\, z^n. \tag{2.168}$$

Using standard z-transform manipulations, equation (2.165) easily transforms into the following recursive equation of the conditional probability generating functions $T(z|k)$:

$$T(z|k) = (\alpha\, z + (1-\alpha)\, z\, B(z))\; T(z|k-1), \tag{2.169}$$

for $k > 1$. In view of the equations (2.166) and (2.167) we further find:

$$T(z|1) = \alpha\, z + (1-\alpha) z B(z). \tag{2.170}$$

The former equation and successive application of equation (2.169) then easily yields an explicit expression for the probability generating function of the effective service time of a customer conditioned on the customer service time and given that the server is available during the slot preceding the effective service time:

$$T(z|k) = \left(\alpha\, z + (1-\alpha)\, z\, B(z)\right)^k. \tag{2.171}$$

We can now average the former expression over all possible customer service times with respect to their probabilities. We then get following expression for the probability generating function of the effective service time of a random customer given that the server was available during the slot preceding the effective service time:

$$T(z) = S\left(\alpha\, z + (1-\alpha)\, z\, B(z)\right). \tag{2.172}$$

As before, taking the appropriate derivatives of the probability generating function (2.172) yields expressions for the various moments of the corresponding random variable. In particular, mean μ_T and variance ${\sigma_T}^2$ of the effective service time for CAI (under the assumption that the effective service time is preceded by an A-slot) are given by

$$\mu_T = \frac{\mu_S}{\sigma} \tag{2.173}$$

and

$${\sigma_T}^2 = \mu_S(1-\alpha)({\sigma_B}^2 + \alpha {\mu_B}^2) + \frac{{\sigma_S}^2}{\sigma^2} \tag{2.174}$$

respectively. The former expressions reveal that the mean (variance) of the effective service time exist when the mean (mean and variance) of the customer service times and the vacation lengths exist for all $0 < \sigma \leq 1$. For $\sigma = 0$, neither mean nor variance of the effective service times exist. In this case, the server is never available.

In general, the n-th order moment of the effective service time exists when moments up to order n of the customer service times and the vacation lengths exist for all $0 < \sigma \leq 1$ and never exist for $\sigma = 0$.

Repeat after interruption with resampling

Let $t(n|k)$ denote the probability that the remaining effective service time of a customer equals n slots, given that it would take k slots (remaining service time) to complete this customer's service under the assumption that there are no vacations. Further let $t(n)$ denote the same probability without conditioning. Recall that one does not need to keep track of the complete service time of the customer in service (the customer's service time is resampled after an interruption) for RAI,wr. That is, RAI,wr is a memoryless operation mode. Conditioning on whether or not a vacation starts during the first slot of the customer's effective service time and on both the length of the vacation and the resampled customer service time if this is the case, we retrieve following recursive equation:

$$\begin{aligned} t(n|k) &= \alpha\, t(n-1|k-1) \\ &\quad + (1-\alpha)\left[b(n-1)s(1) + \sum_{j=1}^{n-2}\sum_{l=2}^{\infty} b(j)\, s(l)\, t(n-j-1|l-1)\right] \end{aligned} \tag{2.175}$$

for $n, k > 1$. Further, a customer's effective service time ends after one slot if and only if his remaining service time equals one slot and if the server is available during the first (and only) effective service time slot, which implies $t(1|1) = \alpha$ and $t(1|k) = 0$ for $k > 1$. Also, given that the customer's remaining service time equals 1 slot, the effective service time equals $n > 1$ slots with probability,

$$t(n|1) = (1-\alpha)\left[b(n-1)s(1) + \sum_{j=1}^{n-2}\sum_{l=2}^{\infty} b(j)\, s(l)\, t(n-j-1|l-1)\right]. \tag{2.176}$$

Let $T(z|k)$ denote the probability generating function that corresponds with $t(n|k)$:

$$T(z|k) = \sum_{n=1}^{\infty} t(n|k)\, z^n \tag{2.177}$$

for $k \geq 1$. Some standard z-transform manipulations then translate equations (2.175) and (2.176) into the following equations for the corresponding probability generating functions:

$$T(z|k) = \alpha\, z\, T(z|k-1) + (1-\alpha)\, z\, B(z)\left[s(1) + \sum_{l=2}^{\infty} s(l)\, T(z|l-1)\right] \tag{2.178}$$

for $k > 1$ and

$$T(z|1) = \alpha\, z + (1-\alpha)\, z\, B(z) \left[s(1) + \sum_{l=2}^{\infty} s(l)\, T(z|l-1) \right]. \tag{2.179}$$

For ease of notation let $\Theta(z)$ denote following probability generating function,

$$\Theta(z) = s(1) + \sum_{l=2}^{\infty} s(l)\, T(z|l-1). \tag{2.180}$$

From equation (2.179) and successive application of equation (2.178) – note that the latter equation is a first order linear recursive equation in k – we obtain the following expression for $T(z|k)$:

$$T(z|k) = (\alpha\, z)^k + \left(1 - (\alpha\, z)^k\right) \frac{(1-\alpha)\, z}{1-\alpha\, z} B(z)\, \Theta(z). \tag{2.181}$$

Plugging the former equation into (2.180) and solving for $\Theta(z)$ now yields,

$$\Theta(z) = \frac{S(\alpha z)(1-\alpha z)}{\alpha z(1-\alpha z) - (1-\alpha) z B(z)\, (\alpha z - S(\alpha z))}. \tag{2.182}$$

By substitution of the former expression in equation (2.181), one finds an expression for the probability generating function of the effective service times conditioned on the customer service time at the beginning of the effective service time:

$$T(z|k) = \frac{(1-\alpha\, z - (1-\alpha) z B(z))(\alpha\, z)^{k+1} + (1-\alpha) z B(z) S(\alpha\, z)}{\alpha\, z(1-\alpha\, z) - (1-\alpha) z B(z)(\alpha\, z - S(\alpha\, z))} \tag{2.183}$$

for $k \geq 1$. Averaging this expression over all possible service times with respect to the customer service time distribution then yields following expression for the probability generating function of the effective service times given that the server is available during the slot preceding the effective service time:

$$T(z) = \frac{S(\alpha z)(1-\alpha z)(\alpha z + (1-\alpha) z B(z))}{\alpha z(1-\alpha z) - (1-\alpha) z B(z)\, (\alpha z - S(\alpha z))}. \tag{2.184}$$

The moment-generating property of probability generating functions then allows us to retrieve various moments of the effective service time for RAI,wr. In particular, the mean μ_T and the variance ${\sigma_T}^2$ (given that an A-slot precedes the effective service time) are given by

$$\mu_T = \alpha \frac{1 - S(\alpha)}{(1-\alpha)\sigma S(\alpha)} \tag{2.185}$$

and

$$\begin{aligned}{\sigma_T}^2 = \alpha \frac{{\sigma_B}^2(1-S(\alpha)) - {\mu_B}^2}{S(\alpha)} + \alpha^2 \frac{1 - 2\,S'(\alpha)\,\sigma\,(1-\alpha)}{\sigma^2(1-\alpha)^2 S(\alpha)^2} \\ - \alpha \frac{1-(1-\alpha)^3\,{\mu_B}^2}{(1-\alpha)^2} + \alpha \frac{1 - 2\,\alpha\mu_B}{(1-\alpha)S(\alpha)} \end{aligned} \tag{2.186}$$

respectively. From the former expressions, one observes that the mean (variance) of the effective service time exists for all $0 < \alpha < 1$ and finite μ_B (μ_B and ${\sigma_B}^2$) as well as for $\alpha = 1$ and finite μ_S (μ_S and ${\sigma_S}^2$). For $\alpha = 0$, the mean (the variance) exists for $S'(0) = s(1) > 0$ and finite μ_B (μ_B and ${\sigma_B}^2$). Note that the available periods all take exactly one slot for $\alpha = 0$. This implies that a customer's service can only complete if the (resampled) customer service time equals one slot. Therefore, moments of the effective service times do not exist if the (resampled) service times never equal one slot ($s(1) = 0$).

In general, the n-th moment of the effective service times exists for all $0 < \alpha < 1$ when the moments up to order n of the vacation distribution exist. Further, for $\alpha = 1$, the n-th moments exist when the moments up to order n of the customer service times exist. For $\alpha = 0$, the n-th moment of the effective service times exists for $s(1) > 0$ when the moments up to order n of the vacation distribution exist.

Again, we observe that in absence of vacations ($\alpha = 1$), existence poses more stringent conditions on the customer service time distribution. This comes from the fact that long service times can be interrupted and resampled to shorter service times in the case of the RAI,wr operation mode.

Repeat after interruption

Recall that we need to keep track of the customer service time in the case of the repeat after interruption mode. This comes from the fact that service has to start all over after an interruption. As for CAI and RAI,wr, we may again deduce a set of recursive equations for the effective service time probabilities (conditioned on the total customer service times and on the remaining customer service time). It is however also possible to retrieve the probability generating function of the effective service time for RAI from the results for RAI,wr. This is how we proceed.

Let $T(z|k)$ denote the probability generating function of the effective service times for RAI, given that the slot preceding the effective service time is an A-slot and given that the customer service time equals k slots. The latter is also the probability generating function of the effective service times for RAI,wr in the particular case that the customer service times are deterministically equal to k slots. Plugging the probability generating function corresponding to deterministic service times $S(z) = z^k$ into

expression (2.184), we get

$$T(z|k) = \frac{(\alpha\, z)^{k-1}\,(1-\alpha\, z)\,(\alpha\, z + (1-\alpha)\, z\, B(z))}{1-\alpha\, z - (1-(\alpha\, z)^{k-1})\,(1-\alpha)\, z\, B(z)} \tag{2.187}$$

for $k > 0$. Averaging over all possible service times with respect to the customer service time distribution then yields the probability generating function of the effective service time given that the server is available during the preceding slot,

$$T(z) = \sum_{k=1}^{\infty} s(k)\, \frac{(\alpha\, z)^{k-1}\,(1-\alpha\, z)\,(\alpha\, z + (1-\alpha)\, z\, B(z))}{1-\alpha\, z - (1-(\alpha\, z)^{k-1})\,(1-\alpha)\, z\, B(z)}. \tag{2.188}$$

Note that this expression is in general not explicit due to the presence of the infinite sum.

The moment-generating property of probability generating functions however allows us again to determine the various moments of the effective service time explicitly. In particular, mean μ_T and variance ${\sigma_T}^2$ of the effective service times in case of RAI operation (given that an A-slot precedes the effective service time) are given by

$$\mu_T = \frac{1}{\sigma}\frac{\alpha}{1-\alpha}\left(S\left(\frac{1}{\alpha}\right) - 1\right) \tag{2.189}$$

and

$$\begin{aligned} {\sigma_T}^2 = & \frac{2\,\alpha^2\, S\left(\frac{1}{\alpha^2}\right) - \alpha^2\, S\left(\frac{1}{\alpha}\right)^2 - \alpha\,\sigma^2}{\sigma^2\,(1-\alpha)^2} \\ & + \frac{\alpha\,\sigma\,(1-2\,\mu_B\,\alpha)\, S\left(\frac{1}{\alpha}\right) - 2\, S'\left(\frac{1}{\alpha}\right)}{\sigma\,(1-\alpha)} \\ & - \alpha\left(\left(1 - S\left(\frac{1}{\alpha}\right)\right){\sigma_B}^2 + {\mu_B}^2\left(\alpha - 1 + S\left(\frac{1}{\alpha}\right)\right)\right) \end{aligned} \tag{2.190}$$

respectively. From the former expressions, we can observe that the mean effective service time in case of RAI operation exists for $1/R_S < \alpha < 1$ and finite mean vacation lengths as well as for $\alpha = 1$ and finite mean customer service time. As before, R_S denotes the radius of convergence of the probability generating function $S(z)$. For $\alpha = R_S$, existence of the mean effective service time depends on the behaviour of S(z) for $z = R_S$. Further, the variance of the effective service time in case of RAI operation exists for $1/\sqrt{R_S} < \alpha < 1$ and finite mean and variance of the vacation lengths as well as for $\alpha = 1$ and finite mean and variance of the customer service times. For $\alpha = 1/\sqrt{R_S}$, existence of the variance of the effective service time depends on the behaviour of $S(z)$ for $z = R_S$.

In general, the n-th moment of the effective service time in case of RAI exists for $1/\sqrt[n]{R_S} < \alpha < 1$ and finite moments up to order n of the vacation length as well as for $\alpha = 1$ and finite moments up to order n of the customer service times. For $\alpha = 1/\sqrt[n]{R_S}$, existence of the n-th moment of the effective service times depend on the behaviour of $S(z)$ for $z = R_S$.

2.4.3 Queue content and customer delay

We may now use the results of the preceding section to establish expressions for the probability generating functions of the queue content and customer delay in a similar way as we did in section 2.3. Here, we follow a different approach. The analysis is based on known results for the exhaustive multiple vacation system which is introduced further. We first focus on (extended) service completion times and their relation to effective service times.

Extended service completion times

Service completion times are defined as the number of slots between the beginning of the slot where the customer receives service for the first time and the end of the slot where the customer leaves the system. One may observe that the server is available during the first and the last slot of the service completion time according to this definition. Therefore, a customer's service completion time does not depend on the state of the server during the slot preceding his service completion time. Further, a customer's extended service completion time is defined as this customer's service completion time if the server is available during the slot following this customer's departure slot or as the sum of this customer's service completion time and the length of the subsequent vacation period if this is not the case. Figure 2.10 illustrates the definitions of the service completion time and the extended service completion time and their relation to the effective service time.

Consider again the effective service time of a customer given that the slot preceding his effective service time is an A-slot. As the slot preceding the customer's effective service time is an A-slot, the effective service time starts with a vacation with probability $1 - \alpha$ or starts with the first slot where the customer receives service with probability α. Therefore, the probability generating function of a customer's effective service time given that the latter is preceded by an A-slot $T(z)$ and the probability generating function of this customer's service completion time $C(z)$ are related as follows:

$$T(z) = (\alpha + (1 - \alpha)B(z))C(z). \tag{2.191}$$

The right-hand side of the former equation also equals the probability generating function of the extended service completion times. This follows from the fact that a cus-

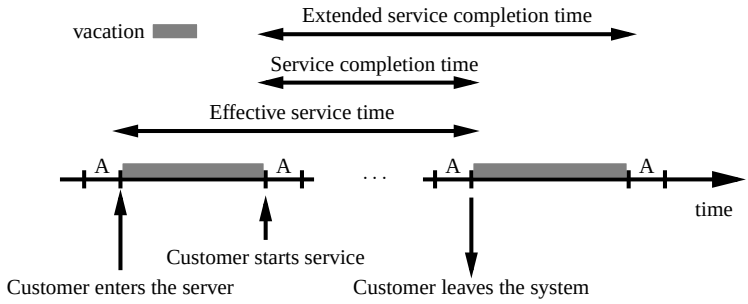


Figure 2.10: Relation between (extended) service completion times and effective service times.

tomer's service completion time is immediately followed by a vacation with probability $(1-\alpha)$ as the last slot of the service completion time is an A-slot by definition. As such, the extended service completion time equals the service completion time with probability α and the sum of the completion time and a vacation with probability $1-\alpha$.

The former observation then yields that $T(z)$ also denotes the probability generating function of the extended service completion times. That is, the probability generating functions of the extended service completion times are given by expressions (2.172), (2.184) and (2.188) for the CAI mode, the RAI,wr mode and the RAI mode respectively.

The exhaustive vacation system

We now concentrate on the discrete-time $Geo^X/G/1$ exhaustive multiple vacation queueing system. The characteristics of this system are as follows.

As always, we assume that time is divided into fixed length slots. During the consecutive slots, customers arrive in the system, are stored in an infinite capacity queue and are served by a single server on a FIFO basis. The numbers of arrivals during the consecutive slots constitute a series of i.i.d. non-negative random variables with common probability generating function $\tilde{E}(z)$. As before, service of customers is synchronised with respect to slot boundaries and the service times of the consecutive customers constitute a series of i.i.d. positive random variables with common probability generating functions $\tilde{S}(z)$. From time to time, the server may leave for a vacation. In particular, the server leaves for a vacation whenever the queue becomes empty. If the queue is still empty upon returning from a vacation, the server immediately leaves for another. The lengths (in slots) of the consecutive vacations also constitute a series of i.i.d. positive random variables. $\tilde{B}(z)$ denotes the corresponding probability generating function.

Takagi [1993] shows that the exhaustive multiple vacation system reaches steady state

if the mean arrival load is less than one, that is, for

$$\mu_{\tilde{E}}\,\mu_{\tilde{S}} < 1. \tag{2.192}$$

Given that the system reaches steady state, the probability generating functions of the queue content at departure epochs in steady state $\tilde{V}_d(z)$ and the probability generating function of the customer delay in steady state $\tilde{D}(z)$ are given by

$$\tilde{V}_d(z) = \frac{1-\mu_{\tilde{E}}\mu_{\tilde{S}}}{\mu_{\tilde{E}}\,\mu_{\tilde{B}}}\frac{\tilde{S}(\tilde{E}(z))}{\tilde{S}(\tilde{E}(z))-z}(1-\tilde{B}(\tilde{E}(z))) \tag{2.193}$$

and

$$\tilde{D}(z) = \frac{1-\mu_{\tilde{E}}\mu_{\tilde{S}}}{\mu_{\tilde{E}}\,\mu_{\tilde{B}}}\frac{1-\tilde{E}(\tilde{S}(z))}{\tilde{E}(\tilde{S}(z))-z}\frac{\tilde{S}(z)}{1-\tilde{S}(z)}(1-\tilde{B}(z)) \tag{2.194}$$

respectively. These results are derived by Takagi [1993] and also show up in section 3.1. There, we investigate the gated-exhaustive vacation system. The latter queueing system encapsulates the exhaustive multiple vacation system.

Bringing everything together

We now combine the extended service completion times and the results for the exhaustive multiple vacation queueing system. In particular, we show how one can retrieve the probability generating functions of queue content and customer delay for the *original system*, that is, the vacation system with generally distributed vacations under consideration.

Consider the exhaustive vacation system characterised by $\tilde{E}(z)$, $\tilde{S}(z)$ and $\tilde{B}(z)$ as follows:

$$\tilde{E}(z) = E(z), \tag{2.195}$$

$$\tilde{S}(z) = T(z), \tag{2.196}$$

$$\tilde{B}(z) = \alpha\, z + (1-\alpha)\, z\, B(z). \tag{2.197}$$

Recall that $E(z)$, $T(z)$ and $B(z)$ are the probability generating functions of the number of arrivals in a slot, of the extended service completion times (or of the effective service times, given that these are preceded by an A-slot) and of the vacations respectively for the original system. Further, $1-\alpha$ denotes the probability that a vacation starts at the end of an A-slot. Clearly, we consider a system with the same arrival process as the original system, with service times corresponding to the extended service completion times of the original system and with vacation lengths equal to one slot or

to one slot more than the length of a vacation in the original system with probability α and $1-\alpha$ respectively. We refer to this system as the *alternative system*.

Substitution of equations (2.195) to (2.197) into equations (2.193) and (2.194) then yields the following expressions for the probability generating functions of queue content at customer departure times and of the customer delay in steady state for the alternative system:

$$\tilde{V}_d(z) = \frac{\sigma(1-\tilde{\rho})}{\mu_E} \frac{T(E(z))}{T(E(z))-z} (1 - \alpha\, E(z) - (1-\alpha)\, E(z)\, B(E(z))), \quad (2.198)$$

$$\tilde{D}(z) = \frac{\sigma(1-\tilde{\rho})}{\mu_E} \frac{1-E(T(z))}{E(T(z))-z} \frac{T(z)}{1-T(z)} (1 - \alpha\, z - (1-\alpha)\, z\, B(z)). \quad (2.199)$$

Here $\tilde{\rho} = \mu_E \mu_T$ denotes the effective load and σ denotes the fraction of available slots as defined in equation (2.164). In view of equation (2.192), the alternative system reaches steady state if $\tilde{\rho} < 1$.

Close observation of the alternative system now reveals that this system operates almost equivalently as the original system under consideration.

First, consider a slot where a random tagged customer receives service for the first time. This customer then occupies the server until his service completion time is finished while a new customer cannot receive service before the tagged customer's extended service completion time is finished. For the alternative system, the tagged customer leaves the system after his extended service completion time and a new customer can receive service during the slot following the customer's departure slot. Therefore, as long as there are customers to serve, customers of the alternative system start receiving service during the same slots as customers of the original system. However, customers of the original system leave after their service completion time whereas customers of the alternative system only leave after their extended service completion time.

Now consider the case when the server finds an empty system after completing a customer's extended service completion time. The server of the original system then starts serving the next customer during the first A-slot following the arrival slot of this customer. Recall that the server of the original system is by definition available during the slot following the extended service completion time. The next A-slot then immediately follows or follows after a vacation and a customer starts service during this A-slot if there are any present. If the system is empty during this A-slot, the next A-slot again either immediately follows or follows after a vacation. This goes on until the server finds a customer waiting. Now consider the alternative system. If the system is empty, the server leaves for a vacation, characterised by the probability generating function $\tilde{B}(z)$, say an E-vacation. If the system is still empty upon returning from a E-vacation, the server leaves immediately for another. It is now easy to identify slots

where the server of the alternative system returns from a E-vacation with the A-slots where customers of the original system can start service. These observations then lead to the conclusion that the systems operate equivalently when the system is empty.

Summarising, we observe that both systems operate equivalently with the following exception. If a vacation immediately starts after the slot where a customer leaves the original system, then the customer only leaves the alternative system after this vacation.

Consider the queue content at departure epochs for the alternative system. Clearly, the latter equals the queue content at departure epochs for the original system if this customer's departure slot is not immediately followed by a vacation. If a vacation immediately follows the departure of a customer of the original system, then the queue content at departure epochs for the alternative system includes all arrivals during this vacation. That is,

$$\tilde{V}_d = V_d + \sum_{j=1}^{B} E^{(j)}. \tag{2.200}$$

Here $\tilde{V}_d$ and V_d denote the queue content at departure epochs of the original and the alternative system, B the length of the vacation and $E^{(j)}$ the number of customer arrivals during the jth slot of this vacation. As the server is available during a customer's departure slot, a vacation follows with probability α. Due to the i.i.d. nature of the arrival process, we easily obtain following relation between the probability generating functions of the queue content for both systems:

$$\tilde{V}_d(z) = V_d(z)\,(\alpha + (1-\alpha)B(E(z))). \tag{2.201}$$

Here $V_d(z)$ denotes the probability generating function of the queue content at departure epochs in steady state of the original system under investigation. Combining the former expression with equation (2.198) then yields,

$$V_d(z) = \frac{\sigma(1-\tilde{\rho})}{\mu_E}\frac{T(E(z))}{T(E(z))-z}\frac{1-\alpha\,E(z)-(1-\alpha)\,E(z)\,B(E(z))}{\alpha+(1-\alpha)B(E(z))}. \tag{2.202}$$

Let $V_r(z)$ denote the probability generating function of the queue content at random slot boundaries in steady state. We then again retrieve the latter using the general relationship (1.63) between the probability generating functions of the queue content at these epochs,

$$V_r(z) = \frac{V_d(z)(z-1)\mu_E}{E(z)-1}. \tag{2.203}$$

Plugging equation (2.202) into the former, we get,

$$V_r(z) = \sigma(1-\tilde{\rho})\frac{T(E(z))(z-1)}{T(E(z))-z}\frac{1-\alpha\,E(z)-(1-\alpha)\,E(z)\,B(E(z))}{(\alpha+(1-\alpha)B(E(z)))(E(z)-1)}. \quad (2.204)$$

If a vacation starts immediately after a customer's service completion, it is included in this customer's delay in the case of the alternative system. As the server is available during the departure slot of the customer, a vacation immediately starts with probability α. Therefore, the probability generating function of the delay for the alternative system $\tilde{D}(z)$ and for the original system $D(z)$ relate as follows:

$$\tilde{D}(z) = D(z)\,(\alpha+(1-\alpha)B(z)). \quad (2.205)$$

Combining the former expression with equation (2.199), we easily find the probability generating function of the customer delay,

$$D(z) = \frac{\sigma(1-\tilde{\rho})}{\mu_E}\frac{1-E(T(z))}{E(T(z))-z}\frac{T(z)}{1-T(z)}\frac{1-\alpha\,z-(1-\alpha)\,z\,B(z)}{\alpha+(1-\alpha)B(z)}. \quad (2.206)$$

Using the moment-generating property of probability generating functions, we can again obtain expressions for the various moments of queue content at random slot boundaries and at departure epochs and for the various moments of the customer delay. In particular the mean queue content at departure epochs and random slot boundaries are given by

$$\begin{aligned}\mu_{V_d} &= \frac{{\sigma_E}^2+{\mu_E}^3\,{\sigma_T}^2-\mu_E\,(1-\tilde{\rho})\,(1-\tilde{\rho}-\mu_E)}{2\,(1-\tilde{\rho})\,\mu_E}\\ &\quad+\frac{\sigma\,\mu_E}{2}\,(1-\alpha)\left(2\,{\mu_B}^2\alpha-{\mu_B}^2-\mu_B+{\sigma_B}^2\right)\end{aligned} \quad (2.207)$$

and

$$\begin{aligned}\mu_{V_r} &= \frac{\tilde{\rho}\,(1-\tilde{\rho})+{\mu_E}^2{\sigma_T}^2+\mu_T{\sigma_E}^2}{2\,(1-\tilde{\rho})}\\ &\quad+\frac{\sigma\,\mu_E}{2}\,(1-\alpha)\left(2\,{\mu_B}^2\alpha-{\mu_B}^2-\mu_B+{\sigma_B}^2\right)\end{aligned} \quad (2.208)$$

respectively whereas the mean customer delay is given by

$$\begin{aligned}\mu_D &= \frac{\tilde{\rho}\,(1-\tilde{\rho})+{\mu_E}^2{\sigma_T}^2+\mu_T{\sigma_E}^2}{2\,\mu_E(1-\tilde{\rho})}\\ &\quad+\frac{\sigma}{2}\,(1-\alpha)\left(2\,{\mu_B}^2\alpha-{\mu_B}^2-\mu_B+{\sigma_B}^2\right).\end{aligned} \quad (2.209)$$

One observes from the former expressions, that these mean values exist when the mean and the variance of the extended service completion times, of the vacation lengths and of the number of customer arrivals in a slot exist. In general, the n-th moment of queue content and customer delay exist whenever the moments up to order $n+1$ of the extended service completion times, the vacation lengths and the number of customer arrivals in a slot exist. For the existence conditions of the moments of the (extended) service completion times, the reader is referred to the preceding subsection.

2.4.4 Idle and busy periods

We now consider idle and busy periods of the system under investigation. The system is idle during a slot when the server is available (an A-slot) and there is no customer in service (the queue is empty at the beginning of the slot). The system is busy whenever the system is not idle. One can easily verify that the consecutive idle and busy periods constitute two independent series of i.i.d. random variables due to both the i.i.d. nature of the arrival process and the memoryless property of the geometrically distributed A-periods.

Consider a random idle-slot. Clearly, the next slot is an idle slot as well if there are no arrivals during this slot and if the server remains available. Therefore the system remains idle with constant probability $\alpha\, E(0)$. This implies that the consecutive idle periods are geometrically distributed. That is, the common probability generating function of the consecutive idle periods $Y_I(z)$ is given by,

$$Y_I(z) = \frac{(1-\alpha\, E(0))\, z}{1-\alpha\, E(0)\, z}. \tag{2.210}$$

We now focus on the joint probability generating function of the consecutive busy periods. Let sub-busy period of a customer denote the number of slots between the beginning of the slot where the customer starts service and the beginning of the slot where – for the first time – the server is available and the queue content is one less than the queue content when this customer starts service. Clearly, the sub-busy period X includes this customer's extended service completion time, as well as all sub-busy periods of arrivals that occur during this extended service completion time, that is,

$$X = T + \sum_{i=1}^{T} \sum_{j=1}^{E^{(i)}} X^{(ij)}. \tag{2.211}$$

Here T denotes the extended service completion time of the customer. Further, $E^{(i)}$ and $X^{(ij)}$ denote the number of arrivals and the sub-busy period of the j-th customer that arrives during the i-th slot of this customer's extended service completion time

respectively. Some standard z-transform manipulations then yield the following functional equation for the probability generating function of the sub-busy periods $X(z)$,

$$X(z) = T(z\,E(X(z))). \tag{2.212}$$

We used the fact that the sub-busy periods constitute an i.i.d. series of random variables. The latter expression defines the probability generating functions of the sub-busy period implicitly and allows us to derive explicit expressions for the moments of the sub-busy periods by evaluation of the appropriate derivatives for $z = 1$.

Let pseudo-busy period denote the number of slots between two consecutive idle-slots. The latter equals 0 when the idle period continues or equals the length of a busy period if this is not the case. This observation then easily leads to following relation between the probability generating function of the pseudo-busy period $Y_{\tilde{B}}(z)$ and the probability generating function of the busy period $Y_B(z)$,

$$Y_{\tilde{B}}(z) = \alpha\,E(0) + (1 - \alpha\,E(0))\,Y_B(z). \tag{2.213}$$

The pseudo-busy period that starts after some idle slot equals the sum of the sub-busy periods of all arrivals during this slot augmented with the length of a vacation and the sum of all sub-busy periods of all arrivals during this vacation if a vacation starts at the end of this slot. The probability generating function of the pseudo-busy period $Y_{\tilde{B}}(z)$ is therefore given by

$$Y_{\tilde{B}}(z) = E(X(z))\,(\alpha + (1-\alpha)B(zE(X(z)))). \tag{2.214}$$

Combining the former two expressions, we easily obtain the joint probability generating function of the consecutive busy periods,

$$Y_B(z) = \frac{E(X(z))\,(\alpha + (1-\alpha)B(zE(X(z))))) - \alpha\,E(0)}{1 - \alpha\,E(0)}. \tag{2.215}$$

The various moments of the idle and busy periods are now easily retrieved with the moment-generating property of probability generating functions. In particular, mean μ_{Y_I} (μ_{Y_B}) and variance ${\sigma_{Y_I}}^2$ (${\sigma_{Y_B}}^2$) of the idle (busy) periods are given by,

$$\mu_{Y_I} = \frac{1}{1 - \alpha\,E(0)}, \tag{2.216}$$

$${\sigma_{Y_I}}^2 = \frac{\alpha\,E(0)}{(1 - \alpha\,E(0))^2} \tag{2.217}$$

and

$$\mu_{Y_B} = \frac{1-\sigma\,(1-\tilde{\rho})}{(1-\tilde{\rho})\,\sigma\,(1-\alpha\,E(0))}, \tag{2.218}$$

$$\begin{aligned}{\sigma_{Y_B}}^2 &= \frac{{\mu_T}^2{\sigma_E}^2+\mu_E{\sigma_T}^2}{\sigma\,(1-\tilde{\rho})^3(1-\alpha\,E(0))} + \frac{(1-\alpha){\sigma_B}^2}{(1-\tilde{\rho})^2(1-\alpha\,E(0))} \\ &\quad + \frac{\left((1-\alpha)\,{\mu_B}^2(1-E(0)) - 2\,\mu_B E(0)\,(1-\alpha)\,\tilde{\rho} - \tilde{\rho}^2 E(0)\right)\alpha}{(1-\tilde{\rho})^2\,(1-\alpha\,E(0))^2}. \end{aligned} \tag{2.219}$$

Let us exclude the (trivial) case that there are neither customer arrivals ($E(0) = 1$) nor vacations ($\alpha = 1$). Clearly, one can observe from the former expressions that under these conditions, the mean and variance of the idle periods always exist. This is also the case for higher order moments of the idle-periods. Existence of the mean busy period requires that the effective load $\tilde{\rho}$ is less than unity as well as the existence of the mean vacation length. Existence of the variance of the busy periods requires again that the effective load $\tilde{\rho}$ is less than unity. Further, existence of the mean and the variance of the vacations, of the number of arrivals in a slot and of the extended service completion times is required. In general, the n-th order moment of the busy period requires that the effective load $\tilde{\rho}$ is less than unity as well as the existence of the moments up to order n of the vacations, of the number of customer arrivals in a slot and of the extended service completion times. The reader is referred to section 2.4.2 where we elaborate on existence conditions for the moments of the effective service times which equal the extended service completion times.

2.4.5 Numerical example

We here concentrate on the characteristics of the mean busy period. Further numerical examples – in the context of priority queueing systems – are the subject of the next section.

Figure 2.11 depicts the mean length of a busy period versus the mean vacation length. We assume that the customer service times share a common shifted symmetrical binomial distribution with mean $\mu_S = 5$. Further, the number of customer arrivals in a slot follows a Poisson distribution and the server is available during half of the slots. That is, σ equals 50% or equivalently, the mean length of the available periods μ_A equals the mean length of the vacation periods μ_B. We here do not need to specify the vacation distribution as the mean busy period does not depend on any vacation process characteristic but the mean lengths of available and blocked periods. We depict the mean length of the busy period for the 3 operation modes under consideration and for μ_E equal to 0.01 and 0.05 as indicated. For CAI, the mean length of the busy period increases for increasing values of the mean length of the vacation periods. This is what one may expect, as vacation periods are included in the busy periods. This is also the case for the RAI and RAI,wr operation modes for large enough values of the

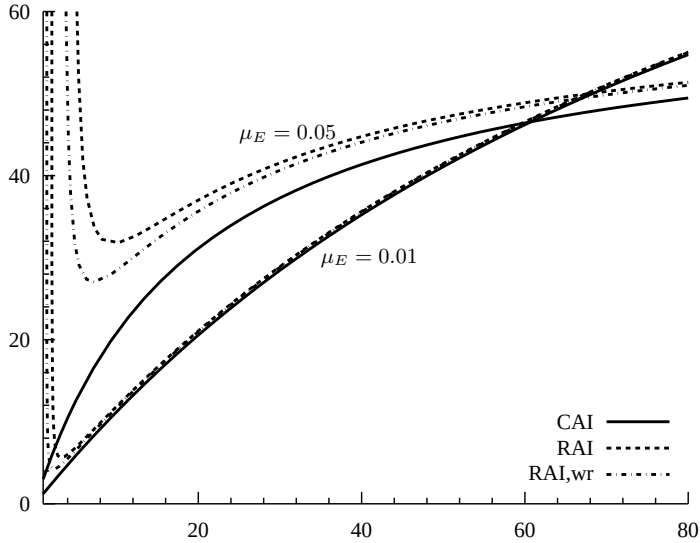


Figure 2.11: The mean busy period vs. the mean vacation period for different values of the mean number of arrivals in a slot. (Poisson arrivals, shifted symmetrical binomially distributed service times with mean $\mu_S = 5$ slots, server availability $\sigma = 0.5$.)

mean length of the vacation periods. Remember that the mean lengths of available periods and vacation periods are equal. Small mean values of the vacation periods therefore imply many service interruptions. As a consequence, the mean (extended) service completion time increases and therefore also the mean busy period increases. Noteworthy is also the fact that for large values of the mean vacation period, the mean busy period is smaller when there are more arrivals. It can easily be shown that for increasing values of the mean length of the vacation periods, the curves tend to,

$$\mu_{Y_B}^{(\infty)} = \frac{1 - \sigma + \mu_E \mu_S}{(\sigma - \mu_E \mu_S)(1 - E(0))}, \tag{2.220}$$

independent of the operation mode under consideration. This comes from the fact that for large vacation periods, there are relatively short busy periods that do not include vacations. This is illustrated in the following figure.

Figure 2.12 depicts the mean length of the busy period versus the mean number of arrivals in a slot. As in the previous figure, we assume that customer service times share a common shifted symmetrical binomial distribution with mean $\mu_S = 5$ and that the number of customer arrivals in a slot follows a Poisson distribution. The server is again available for half of the time ($\sigma = 0.5$). We depict the mean length of the busy periods for all 3 operation modes and for the cases that the mean vacation length μ_B equals 10 and 100 slots. For zero load ($\mu_E = 0$), the mean busy period equals the mean vacation length as in this case busy and vacation periods correspond. For

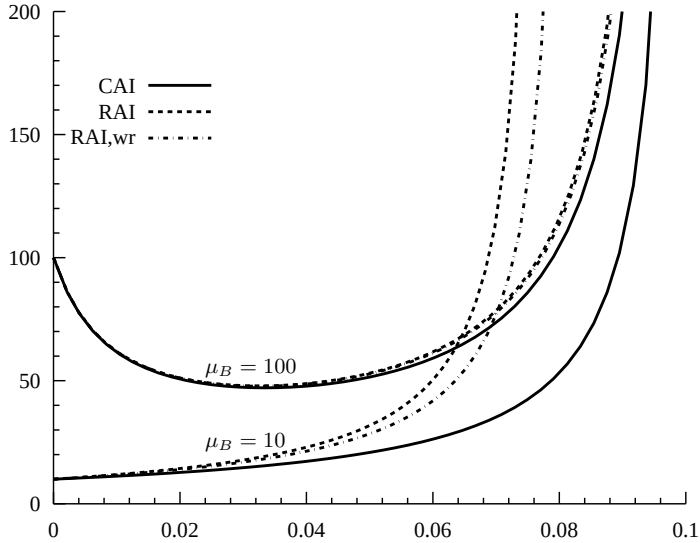


Figure 2.12: The mean busy period vs. the mean number of customer arrivals in a slot μ_E for different values of the mean vacation length. (Poisson arrivals, shifted symmetrical binomially distributed service times with mean $\mu_S = 5$ slots, server availability $\sigma = 0.5$.)

$\mu_B = 10$, increasing load implies an increase of the mean busy period as expected. This is however not the case for $\mu_B = 100$. Larger values of the mean vacation length imply larger values of the mean available period. For sufficiently low arrival load, the probability that a busy period starts and ends during the same available period is high. The mean length of this type of busy periods is a lot shorter than the mean length of the vacation periods. For increasing load, the number of this type of the busy periods first increases (more arrivals) and then decreases (longer busy periods, increasing chance that a vacation starts). As a consequence, the mean length of the busy periods first decreases and then increases for increasing values of the mean number of customer arrivals during a slot. Further, for both $\mu_B = 10$ and $\mu_B = 100$, the mean busy period is shorter for CAI compared to RAI,wr. In turn the mean busy period for RAI,wr is shorter in comparison with the RAI mode. This does not come as a surprise as a part of the service gets lost for RAI and RAI,wr due to service interruptions.

2.5 Preemptive priority queueing systems

In this section we consider multi-class queueing systems with a preemptive priority scheduling discipline. We show that the vacation model presented in the preceding section can be used to assess performance of this type of queueing systems. Part of this section follows the lines of our contribution [Fiems et al., 2003]. We first survey

some literature on the performance evaluation of queueing systems with a priority scheduling discipline. In particular, we concentrate on discrete-time queueing models with a priority scheduling discipline. For continuous-time models with priorities, the reader is referred to the monographs of Kleinrock [1976] and Takagi [1991] and the references therein.

2.5.1 Priority models

Priority scheduling disciplines come in two basic flavours: the preemptive priority and the non-preemptive priority scheduling discipline. For a preemptive priority scheduling discipline, service of a customer is interrupted when a customer with higher priority arrives in the system during this customer's service. The interrupted customer only gets hold of the server again when there are no more higher-priority customers present in the system. After such an interruption, the customer either resumes his service (*preemptive resume*), repeats his service (*preemptive repeat*) or repeats his service with a new service time sample (*preemptive repeat with resampling*). On the other hand, for non-preemptive scheduling, a customer's service is never interrupted. Upon departure of a customer or when customers arrive in an empty system, the server selects a customer for service from the class with the highest priority with waiting customers.

Different authors consider discrete-time priority queueing systems with single-slot service times. Takine et al. [1994] consider a two-class priority system with Markov-modulated high-priority arrivals and uncorrelated low-priority arrivals. Laevens and Bruneel [1998] investigate a multi-server queueing system with priorities whereas Walraevens et al. [2003d] focus on an output queueing switch with two priority classes. Notice that in the case of single-slot service times, there is no need to differentiate between non-preemptive and preemptive – resume, repeat or repeat with resampling – scheduling disciplines as service is never interrupted.

Systems with multiple-slot service times and a non-preemptive priority scheduling discipline are investigated by a.o. Rubin and Tsai [1989], Walraevens et al. [2002, 2003b] and Takahashi et al. [1999]. Rubin and Tsai [1989] consider the customer delay in a multi-class non-preemptive queueing system with uncorrelated arrivals. Walraevens et al. [2002] concentrate on non-preemptive systems with two priority classes. Although there is no correlation in the arrival process from slot to slot, these authors allow correlation between customer arrivals of the different classes within a slot. The same authors [2003b] also consider a non-preemptive queueing system with 3 classes. Finally, Takahashi et al. [1999] consider a two-class non-preemptive queueing system where the low-priority customers do not wait for service but come back after some time. I.e., these authors consider a retrial queueing system with non-preemptive priorities.

Preemptive priority systems are considered by a.o. Rubin and Tsai [1989], Lee [2001], Lee and Lee [2003] and Walraevens et al. [2003a,c]. The preemptive resume priority

discipline [Rubin and Tsai, 1989, Lee, 2001, Walraevens et al., 2003c] as well as the preemptive repeat with resampling priority discipline [Lee and Lee, 2003, Walraevens et al., 2003a] are investigated. To our knowledge, discrete-time queueing systems with a preemptive repeat scheduling discipline have not been considered before.

2.5.2 Queueing model and analysis

In this section, we consider a discrete-time multi-class preemptive priority system. The numbers of arrivals of the different classes during the consecutive slots as well as their service times are assumed to constitute series of i.i.d. random variables. Further, the arrivals and service times in each class are assumed to be independent from the latter quantities in other classes. We investigate the preemptive resume, the preemptive repeat and the preemptive repeat with resampling scheduling disciplines. The scheduling discipline can be chosen independently from class to class allowing maximal flexibility. Further, customers of the same class are served according to a FIFO scheduling discipline. The mean number of class i arrivals in a slot is denoted by μ_{E_i} and their mean service time by μ_{S_i}. As such, the class i load is given by $\rho_i = \mu_{E_i}\mu_{S_i}$. We assume that class 1 corresponds to the class with the highest priority.

As preemptiveness implies that lower-priority classes do not influence performance of higher-priority classes, class 1 performance is easily assessed by means of a standard $Geo^X/G/1$ queueing model (see section 1.2). Consecutive busy and idle periods of a discrete-time $Geo^X/G/1$ queueing system constitute two independent series of i.i.d. random variables with general and geometrical distributions respectively. This implies that class 2 customers perceive server availability as an on/off process with generally distributed B-periods and geometrically distributed A-periods. Therefore, we may assess performance of class 2 customers using the queueing model presented in the preceding section. The operation modes CAI, RAI,wr and RAI correspond to the preemptive resume, the preemptive repeat with resampling and the preemptive repeat without resampling priority scheduling disciplines respectively.

A similar approach is also possible for lower-priority classes if more than two classes are involved. That is, the vacation model of the preceding section can be used to assess performance of any class. This follows from the fact that consecutive idle and busy periods of this model – remember that busy periods include the slots during which the server is on vacation – constitute two independent series of i.i.d. random variables with general and shifted geometrical distributions respectively as well. The vacation (available) periods perceived by class k correspond to the busy (idle) periods of class $k-1$. Therefore, performance assessment of class k customers requires the iterative calculation of the probability generating functions of idle and busy periods up to class $k-1$. For each class, we may select one of the three operation modes (CAI, RAI,wr or RAI) independently of the other classes. That is, we can select one of the three preemptive priority scheduling disciplines for each class.

2.5.3 Approximate analysis

Although we can model a preemptive multi-class priority system exactly, we may also look into approximate results. In particular we investigate here whether it is possible to investigate preemptive priority systems by means of the Bernoulli model and the 2-state Markovian model as presented in sections 2.1 and 2.3 respectively. We here limit our investigation to the case of two customer classes.

Recall that the Bernoulli vacation process is completely characterised by the probability σ that the server is available during a random slot. Class 2 customers perceive an available server when there are no class 1 customers present. Further, class 1 customers occupy the server for a fraction ρ_1 of the slots. To assess performance of class 2 customers, we therefore assume,

$$\sigma = 1 - \rho_1. \tag{2.221}$$

Notice that in case of the Bernoulli model, correlation in the perceived vacation model is completely neglected.

As opposed to the Bernoulli model, the 2-state Markovian vacation process allows us to specify two independent transition probabilities, or, equivalently, the mean length of an A-period and the mean length of a B-period. To assess performance of class 2 customers we therefore map the mean length of an A-period μ_A and of a B-period μ_B to the mean length of a class 1 idle period and of a class 1 busy period respectively. That is,

$$\mu_A = \frac{1}{1 - e_1(0)}, \tag{2.222}$$

$$\mu_B = \frac{1}{1 - e_1(0)} \frac{\rho_1}{1 - \rho_1}. \tag{2.223}$$

Expressions for the mean class 1 idle and busy period follow from plugging $\alpha = \sigma = 1$ (no vacations) into expressions (2.216) and (2.218). Here $e_1(0)$ denotes the probability that there are no class 1 arrivals in a slot. As opposed to the Bernoulli model, the 2-state Markovian model allows us to take some of the correlation in the perceived vacation process into account.

We now investigate the accuracy of these approximations by means of some numerical examples. Figures 2.13 and 2.14 depict the mean and the variance of the class 2 queue content – the number of class 2 customers in the system – at random slot boundaries versus the total system load respectively for a two-class priority system. The total system load is the sum of the loads of the different classes. We here assume that 20% of the total load is class 1 load. Further, the number of customer arrivals in a slot of either class follows a Bernoulli distribution whereas the service times of both class 1 and class 2 customers follow a shifted symmetric binomial distribution with

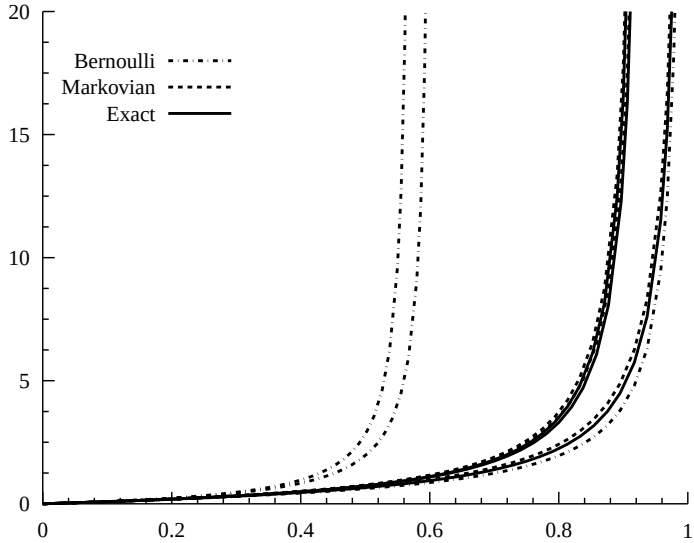


Figure 2.13: The mean class 2 queue content vs. the system's load for – from left to right – the preemptive repeat, the preemptive repeat with resampling and the preemptive resume disciplines. (2 priority classes; class 1: Bernoulli arrivals, shifted symmetrical binomial service times with mean $\mu_{S_1} = 10$ slots, 20% of the total load; class 2: Bernoulli arrivals, shifted symmetrical binomial service times with mean $\mu_{S_2} = 10$ slots, 80% of the total load.)

a mean value of 10 slots. The reader is referred to appendix A for details on these distributions. For each of the approximations and for the exact model, we depict the mean class 2 queue content in case of – from left to right – the preemptive repeat, the preemptive repeat with resampling and the preemptive resume priority disciplines.

First consider the mean queue content (figure 2.13). For the preemptive resume discipline, both approximations are fairly accurate. This is also the case for the 2-state Markov approximation for the preemptive repeat and the preemptive repeat with resampling priority disciplines. The Bernoulli model however cannot be used to assess performance for the preemptive repeat and the preemptive repeat with resampling priority disciplines. As service interruptions imply lost service time, performance in these cases heavily depends on the characteristics of the A-periods. The Bernoulli model does not allow us to capture these characteristics sufficiently accurately. Regarding the variance of the class 2 queue content (figure 2.14), we again observe that the Markov model yields fairly accurate results for all operation modes. The Bernoulli model on the other hand is completely inaccurate.

Summarising, we find that the 2-state Markov can predict performance of class 2 customers fairly accurately. This is definitely not the case for the Bernoulli model. One cannot completely neglect correlation in the perceived vacation model.

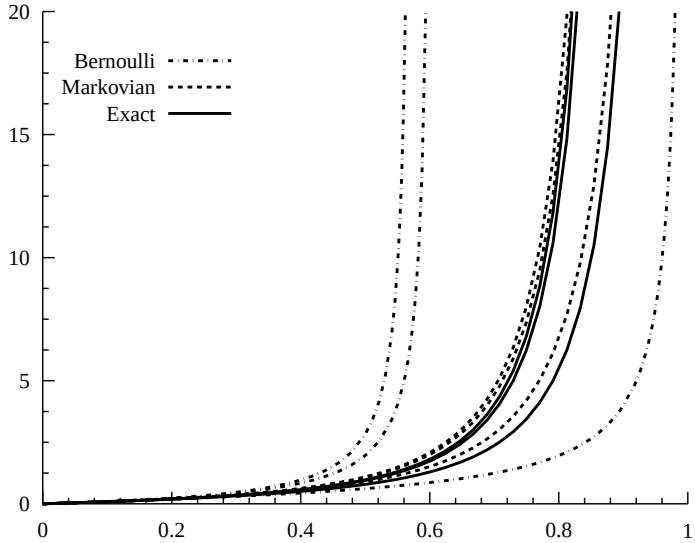


Figure 2.14: The variance of the class 2 queue content vs. the system's load for – from left to right – the preemptive repeat, the preemptive repeat with resampling and the preemptive resume disciplines. (2 priority classes; class 1: Bernoulli arrivals, shifted symmetrical binomial service times with mean $\mu_{S_1} = 10$ slots, 20% of the total load; class 2: Bernoulli arrivals, shifted symmetrical binomial service times with mean $\mu_{S_2} = 10$ slots, 80% of the total load.)

2.5.4 Numerical examples

By means of some numerical examples, we now investigate the performance of multi-class preemptive priority systems. We here focus on exact results. We assume that at most one customer of each class can arrive in the system during a slot for all numerical examples. That is, for each class, the numbers of the customer arrivals of this class during the consecutive slots are modelled by means of series of independent Bernoulli distributed random variables.

Figures 2.15 and 2.16 depict the mean delay of customers of the different classes versus the total system load for a 4-class preemptive priority system. Of the total system load, a fraction of 10% is class 1 load, a fraction of 20% is class 2 load, a fraction of 30% is class 3 load and a fraction of 40% is class 4 load. In figure 2.15, customer service times of all classes share a shifted geometrical distribution with a mean of 10 slots. We consider the preemptive resume and the preemptive repeat priority disciplines for all classes. Note that in this case, due to the lack of memory of the geometrical distribution, preemptive resume and preemptive repeat with resampling disciplines operate equivalently. In figure 2.16, customer service times are deterministically equal to 10 slots. We again consider the preemptive resume and the preemptive repeat priority scheduling disciplines for all classes. As resampling of a

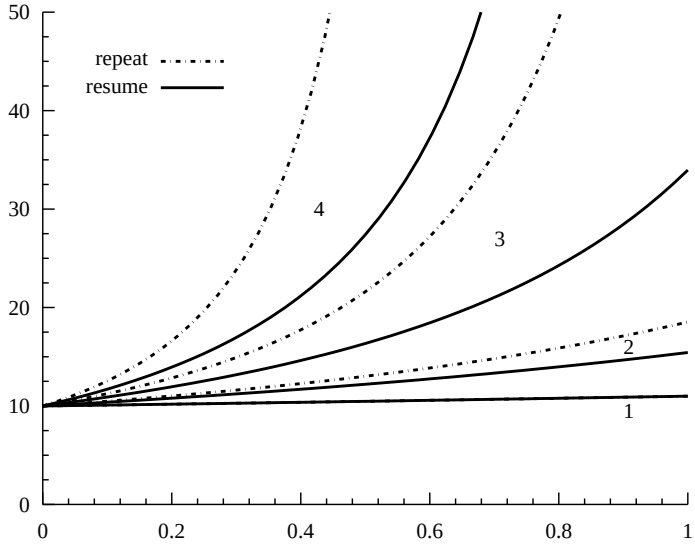


Figure 2.15: Mean delay vs. total load. (Bernoulli arrivals; 4 priority classes: 10% class 1 load, 20% class 2 load, 30% class 3 load and 40% class 4 load; shifted geometrical service times with mean $\mu_S = 10$ slots for all classes.)

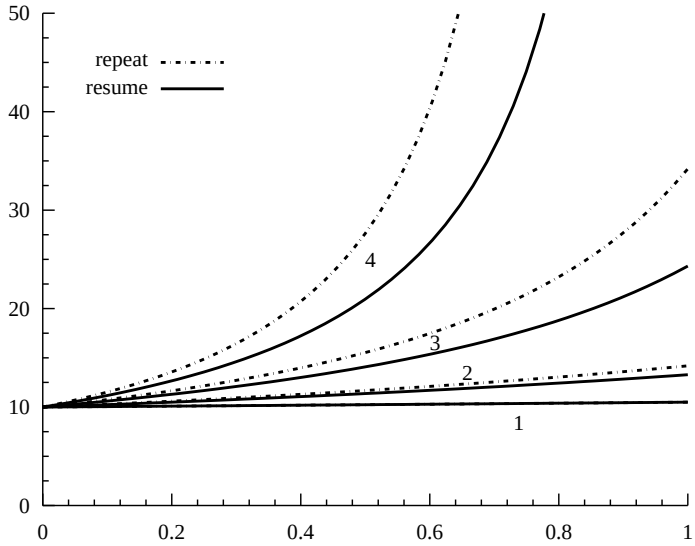


Figure 2.16: Mean delay vs. total load. (Bernoulli arrivals; 4 priority classes: 10% class 1 load, 20% class 2 load, 30% class 3 load and 40% class 4 load; deterministic service times with mean $\mu_S = 10$ slots for all classes.)

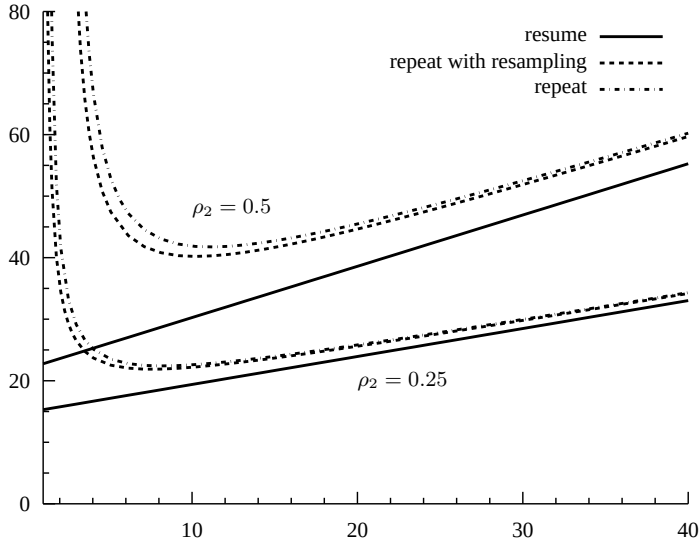


Figure 2.17: Mean class 2 customer delay vs. mean class 1 service time for different values of the class 2 load ρ_2. (2 priority classes; class 1: Bernoulli arrivals, shifted geometrical service times, class 1 load $\rho_1 = 0.2$; class 2: Bernoulli arrivals, shifted symmetrical binomially distributed service times with mean $\mu_{S_2} = 10$.)

deterministic random variable yields by definition the same value, preemptive repeat and preemptive repeat with resampling disciplines operate equivalently. As expected, the preemptive priority discipline allows us to boost performance of high-priority customers at the expense of deteriorating performance of lower-priority customers. Further, the preemptive resume discipline outperforms the preemptive repeat discipline as service repetitions cause performance degradation.

We now focus on two-class priority systems. In figures 2.17 and 2.18 we investigate the influence of the mean class 1 customer service times on class 2 performance, while we keep the class 1 load fixed. Figure 2.17 depicts the mean class 2 customer delay whereas figure 2.18 depicts the corresponding variance. Class 2 customer service times follow a shifted symmetrical binomial distribution with a mean value of 10 slots whereas class 1 service times follow a shifted geometrical distribution. The class 1 load equals 0.2 for all curves whereas the class 2 load equals either 0.25 or 0.5 as indicated. For the preemptive resume discipline, both mean and variance increase for increasing mean values of the class 1 customer service times. This comes from the fact that for fixed class 1 load, longer class 2 queues are built up during longer class 1 busy periods which are implied by longer class 1 service times. That is, the class 2 customers perceive a more bursty vacation process which implies that performance deteriorates. However, whenever service repetitions are necessary, additional burstiness in the (by the class 2 customers) perceived vacation process not only implies

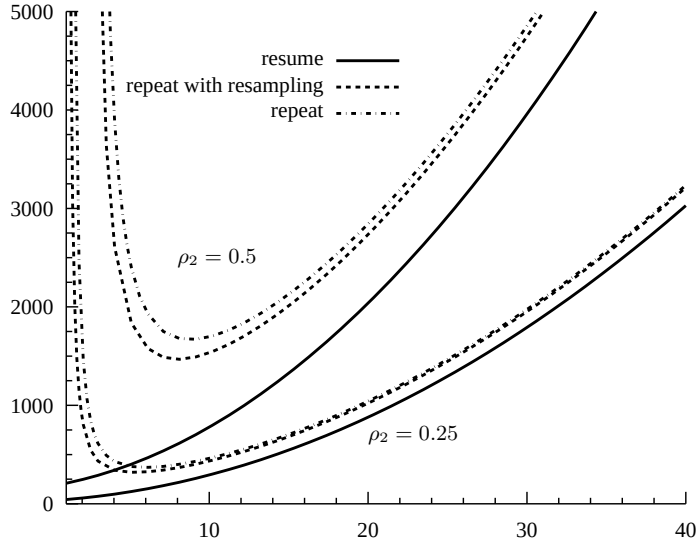


Figure 2.18: Variance of class 2 customer delay vs. mean class 1 service time for different values of the class 2 load ρ_2. (2 priority classes; class 1: Bernoulli arrivals, shifted geometrical service times, class 1 load $\rho_1 = 0.2$; class 2: Bernoulli arrivals, shifted symmetrical binomially distributed service times with mean $\mu_{S_2} = 10$.)

longer build-up periods but also less service interruptions (and repetitions). This follows from the fact that increasing burstiness implies longer available periods. For long class 1 customer service times, class 2 service interruptions are already scarce and the former effect dominates whereas the latter effect may dominate for short class 1 customer service times. These effects explain the shapes of the curves for the preemptive repeat and the preemptive repeat with resampling priority disciplines.

In figure 2.19, we finally consider the upper bound for the fraction of the total load which can be served with priority under mean delay constraints versus the total load. Service times of both class 1 and class 2 customers obey a shifted symmetrical binomial distribution with mean $\mu_{S_1} = \mu_{S_2} = 10$ slots.

Class 1 mean delay constraints limit the fraction that can be sent with priority as more class 1 customer arrivals imply longer class 1 delays. For lower values of the total load, all customers can be served with priority. In this case, the system operates as a FIFO queueing system. For increasing values of the load, a FIFO discipline no longer suffices to meet the mean class 1 delay constraint. To guarantee this constraint, only a fraction of the load can be served with priority. Consider now mean class 2 delay constraints. Here we have to differentiate between the different preemptive priority disciplines under consideration. For low values of the total load, all customers can be served with priority. For higher values of the load however, class 2 delay constraints limit the fraction of class 1 load. Performance of class 2 customers deteriorates due to

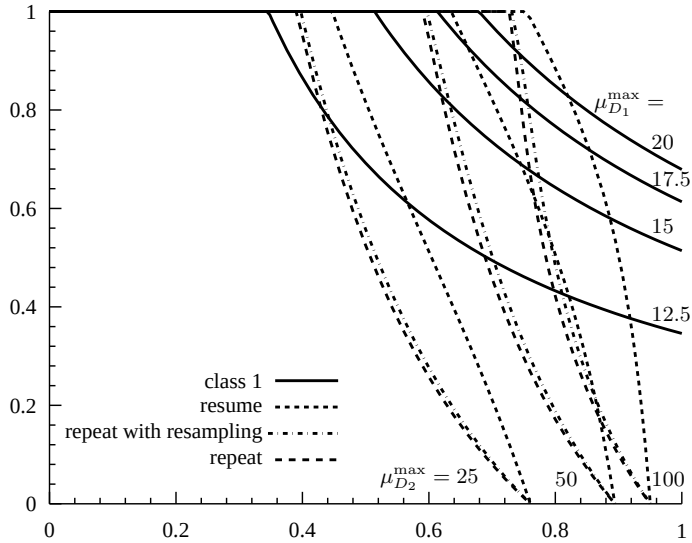


Figure 2.19: Maximal class 1 fraction vs. the load under delay constraints. (2 priority classes; For both classes: Bernoulli arrivals, shifted symmetrical binomially distributed service times with mean $\mu_{S_1} = \mu_{S_2} = 10$ slots.)

the presence of class 1 load and therefore the latter should be limited. For even higher values of the load, delay constraints cannot be met. That is, even if all customers are served without priority – the system then operates as a FIFO discipline – mean class 2 delay is higher than the constraint. Therefore, the maximal total load for which class 2 delay constraints can be met corresponds to the maximal load in a FIFO queue for which this constraint is met. Clearly, this value does not depend on the preemptive priority discipline under consideration.

2.6 Other operation modes

So far, we considered three different operation modes to cope with interrupted service. After the interruption, service was either resumed or repeated. In the latter case we investigated the case that the service time remains the same and the case that the service time is resampled after the interruption. This section is concerned with some other operation modes. In particular, we focus on *delayed* and *partial* operation modes. We hereby limit ourselves to the $Geo^X/G/1$ model with Bernoulli vacations. The remainder of this section closely follows the lines of our contribution [Fiems et al., 2002c].

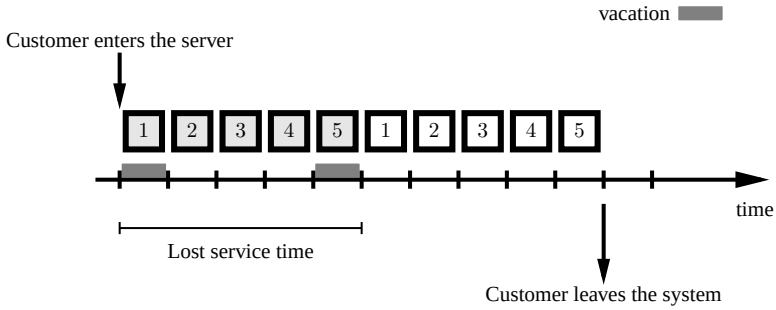


Figure 2.20: Effective service time for the delayed repeat after interruption mode.

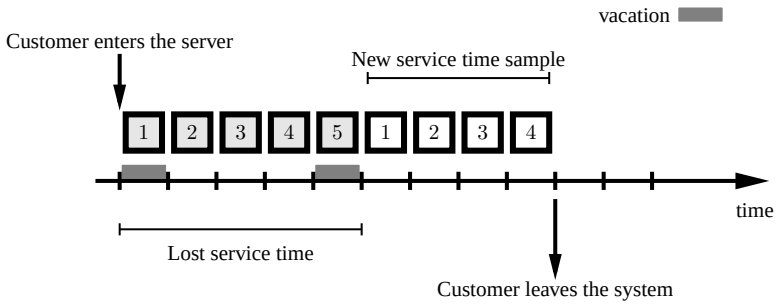


Figure 2.21: Effective service time for the delayed repeat after interruption with resampling mode.

2.6.1 Delayed operation modes

Delayed operation modes take into account that vacations may not be detected immediately. For these operation modes, service continues during vacations. For the *delayed repeat after interruption* mode (d-RAI) and the *delayed repeat after interruption with resampling* mode (d-RAI,wr), service is repeated until the customer receives service without vacations during his service time. Repetitions require equal service time for the d-RAI mode whereas service time is resampled for every repetition in the case of the d-RAI,wr mode.

Recall that the effective service time of a customer is defined as the number of slots between the beginning of the slot where this customer enters the server and the end of the slot where this customer leaves the system. For the delayed operation modes, the effective service times therefore include all necessary repetitions. Figures 2.20 and 2.21 illustrate the definition of the effective service times for the d-RAI and the d-RAI,wr operation modes respectively. This customer's (original) customer service equals 5 slots and is resampled to 4 slots in case of the d-RAI,wr operation mode.

Remember that the consecutive effective service times in case of the Bernoulli vacation model constitute a series of i.i.d. random variables. This implies that the derivation of the probability generating functions of the effective service times reduces the system analysis to the analysis of an equivalent system without vacations but with service times given by the effective service times. That is, the system analysis reduces to the analysis of the $Geo^X/G/1$ queueing system. We here focus on the effective service times. The reader is referred to sections 1.2 and 2.1 for the queueing analysis of the $Geo^X/G/1$ system.

Delayed RAI with resampling

We first consider the delayed repeat after interruption with resampling service mode. The effective service time T of a random customer equals the (original) service time S of this customer if the server remains available during this service time. If this is not the case, the effective service time equals the sum of this customer's service time and the remaining effective service time $\hat{T}$ thereafter.

$$T = \begin{cases} S & \text{no vacations during service,} \\ S + \hat{T} & \text{vacations during service.} \end{cases} \tag{2.224}$$

From a system's point of view, there is no difference between serving the same customer once more and serving a new customer. This follows from the Bernoulli nature of the vacation process and the fact that customer service times are resampled. As a consequence, the effective service time T and the remaining effective service time $\hat{T}$ share a common probability generating function. One easily verifies that the probability that there are no vacations during a service time of S slots equals σ^S. As before, σ denotes the probability that the server is available during a slot. Some standard z-transform manipulations and equation (2.224) then yield the following expression for the probability generating function $T(z)$ of a random customer's effective service time:

$$T(z) = S(\sigma z) + (S(z) - S(\sigma z))T(z). \tag{2.225}$$

Here $S(z)$ denotes the probability generating function of the customer service times. Solving for $T(z)$ then yields,

$$T(z) = \frac{S(\sigma\, z)}{1 - S(z) + S(\sigma\, z)} \tag{2.226}$$

The moment-generating property of probability generating functions again allows us to obtain moments of the effective service times in case of the d-RAI,wr interruption

operation mode and the mean μ_T and variance ${\sigma_T}^2$ are in particular given by

$$\mu_T = \frac{\mu_S}{S(\sigma)} \tag{2.227}$$

and

$${\sigma_T}^2 = \frac{S(\sigma)\left({\sigma_S}^2 + {\mu_S}^2\right) + {\mu_S}^2 - 2\,\mu_S\,\sigma\,S'(\sigma)}{S(\sigma)^2} \tag{2.228}$$

respectively. From the former expressions one easily observes that for $0 < \sigma \leq 1$, mean (variance) of the effective service times exist whenever the mean (mean and variance) of the underlying service time distribution exist.

More general, the n-th moment of the effective service time can be expressed in terms of moments of the underlying service time distribution up to order n and of the generating function and its derivatives evaluated in σ. Therefore, the n-th moment exists whenever the moments up to order n of the underlying service time distribution exist.

Delayed RAI

For the delayed RAI (without resampling) operation mode, we first consider the effective service times conditioned on the customer service time. Let $T(z|k)$ denote the probability generating function of a customer's effective service time given that his service time equals k slots. The latter may be derived by conditioning on the number of necessary repetitions. However, d-RAI and d-RAI,wr operate equivalently for deterministic customer service times. We therefore retrieve $T(z|k)$ by plugging the probability generating function $S(z) = z^k$ of deterministic customer service times into the expression (2.226) of the probability generating function of the effective service times for d-RAI,wr. We find,

$$T(z|k) = \frac{(\sigma\,z)^k}{1 - z^k + (\sigma\,z)^k}. \tag{2.229}$$

Averaging over all possible service times with respect to the service time probabilities $s(k)$ ($k > 0$) then yields the probability generating function $T(z)$ of the effective service times in case of d-RAI:

$$T(z) = \sum_{k=1}^{\infty} s(k)\,\frac{(\sigma\,z)^k}{1 - z^k + (\sigma\,z)^k}. \tag{2.230}$$

As for RAI, we may again retrieve explicit expressions for the various moments des-

pite the infinite sum. In particular mean and variance are given by

$$\mu_T = \frac{S'\left(\frac{1}{\sigma}\right)}{\sigma} \tag{2.231}$$

and

$$\sigma_T{}^2 = \frac{2\,S''\left(\frac{1}{\sigma^2}\right) - S'\left(\frac{1}{\sigma}\right)^2 \sigma^2 - S''\left(\frac{1}{\sigma}\right)\sigma^2 - S'\left(\frac{1}{\sigma}\right)\sigma^3 + 2\,S'\left(\frac{1}{\sigma^2}\right)\sigma^2}{\sigma^4} \tag{2.232}$$

respectively. As for RAI, higher order moments may be expressed in terms of the generating function of the service times and its derivatives evaluated in negative integer powers of σ. In particular, existence of the n-th moment is guaranteed whenever,

$$R_S^{-\frac{1}{n}} < \sigma \leq 1. \tag{2.233}$$

Here R_S denotes the radius of convergence of the probability generating function $S(z)$. Existence of the moments for $\sigma = R_S^{-\frac{1}{n}}$ depends on the behaviour of the probability generating function and its derivatives on its radius of convergence. Notice that existence of moments for some value $\sigma < 1$ requires at least $R_S > 1$ and therefore all moments of the service time distribution exist.

2.6.2 Partial interruption modes

The former operation modes with repetitions (RAI, RAI,wr, d-RAI, d-RAI,wr) always considered the case where service time was repeated as a whole. We now investigate operation modes where only a part of the service time is repeated. Therefore we assume that a customer's service time contains service parts and do not apply the former operation modes on the complete service time but on these service parts instead.

We assume that the numbers of service parts required by the consecutive customers constitute a series of i.i.d. positive random variables with common probability generating function $S_p(z)$. Also, the service times required by the consecutive parts constitute a series of i.i.d. positive random variables. Their common probability mass function and probability generating function are denoted by $l(n)$ ($n \geq 1$) and $L(z)$. The complete customer service time equals the sum of the service times of the parts. Due to the independence assumptions, the probability generating function of the customer service times $S(z)$ is given by,

$$S(z) = S_p(L(z)). \tag{2.234}$$

The partial interruption modes under consideration are in particular the *partial repeat after interruption* mode (p-RAI), the *partial repeat after interruption with resampling*

mode (p-RAI,wr), the *delayed partial repeat after interruption* mode (dp-RAI) and the *delayed partial repeat after interruption with resampling* mode (dp-RAI,wr). Note that all these operation modes reduce to CAI in case the parts deterministically require a single-slot service time. In this case, we find $L(z) = z$ and $S(z) = S_p(z)$. Further, the operation modes reduce to their non-partial equivalents if each customer's service time deterministically contains one part. In this case, we find $S_p(z) = z$ and $S(z) = L(z)$.

The independence of the vacation process implies that the effective service times of the consecutive service parts constitute a series of i.i.d. random variables. The effective service time of a service part starts at the beginning of the slot where "the part enters the server" and ends at the end of the slot where the service of the part is completed. One obtains the common probability generating functions of the effective service times of the parts for the different operation modes by replacing the generating function of the customer service times $S(z)$ by the generating function of the service times of the parts $L(z)$. One may retrieve these generating functions from equations (2.14), (2.20), (2.226) and (2.230) for p-RAI, p-RAI,wr, dp-RAI,wr and dp-RAI respectively. As the effective service times of the parts constitute a series of i.i.d. random variables, the probability generating function of the customer's total effective service time – that is, the sum of the effective service times of the parts – is given by,

$$T(z) = S_p(\tilde{T}(z)). \tag{2.235}$$

Here $\tilde{T}(z)$ denotes the probability generating function of the effective service time of a part for one of the operation modes under consideration. Explicit expressions for the probability generating functions of the effective service times for the different partial modes are displayed in table 2.1.

Existence of the n-th moment of a customer's effective service time for a partial operation mode requires existence of the moments of the number of parts constituting a customer's service time and of the moments of the effective service times of the parts up to order n. In the case of the p-RAI and dp-RAI operation modes, the latter moments exist whenever $\sigma > R_L^{-\frac{1}{n}}$. Here, R_L denotes the radius of convergence of $L(z)$. The moments of the effective service times up to order n always exist for p-RAI,wr and require the existence of the moments of the service times of the parts up to order n for d-RAI,wr.

The moment-generating property of probability generating functions then allows to retrieve various moments of the effective service times for the different partial operation modes. In particular, the mean and the variance are displayed in tables 2.2 and 2.3 respectively. Here μ_{S_p} (${\sigma_{S_p}}^2$) and μ_L (${\sigma_L}^2$) denote the mean (variance) of the service times of the parts and the mean (variance) of the number of parts constituting a customer's service time respectively.

mode	pgf of the effective service times
p-RAI	$S_p\left(\sum_{k=1}^{\infty} l(k)\frac{(\sigma z)^k(1-\sigma z)}{1-z+(1-\sigma)z(\sigma z)^k}\right)$
p-RAI,wr	$S_p\left(\frac{L(\sigma z)(1-\sigma z)}{1-z+(1-\sigma)\,z\,L(\sigma z)}\right)$
dp-RAI	$S_p\left(\sum_{k=1}^{\infty} l(k)\frac{(\sigma z)^k}{1-z^k+(\sigma z)^k}\right)$
dp-RAI,wr	$S_p\left(\frac{L(\sigma z)}{1-L(z)+L(\sigma z)}\right)$

Table 2.1: Probability generating functions of the effective service times for the partial operation modes

2.6.3 Some numerical examples

Remember that the efficiency ϵ of an operation mode is defined as the mean effective service time when no service is lost divided by the mean effective service time of the operation mode under consideration. That is,

$$\epsilon = \frac{\mu_S}{\sigma\,\mu_T}. \tag{2.236}$$

From the definition above, it is clear that the efficiency depends on the server availability σ as well as on the customer service time distribution.

Figures 2.22 and 2.23 depict the efficiency for the d-RAI and the d-RAI,wr operation modes versus the mean length of an available period μ_A. We consider the cases where the customer service times share a deterministic, a shifted Poisson and a shifted geometric distribution. The reader is referred to appendix A for details on these distributions. The mean customer service time equals 5 slots and 10 slots for figures 2.22 and 2.23 respectively. For deterministically distributed customer service times, the efficiency for d-RAI equals the efficiency for d-RAI,wr as these modes operate equivalently when customer service times are deterministically distributed. Further, given the same (non-deterministic) customer service time distribution, d-RAI,wr outperforms d-RAI. It can be shown that given any customer service distribution, efficiency for d-RAI,wr is equal to or larger than the efficiency for d-RAI (as for the non-delayed operation modes, the proof is based on Jensen's inequality). Recall that for the non-delayed operation modes, we showed that, given the mean customer service time, the efficiency of RAI,wr is at least as good as the efficiency for RAI. This

mode	mean effective service time
p-RAI	$\dfrac{\mu_{S_p}\left(L\left(\frac{1}{\sigma}\right)-1\right)}{1-\sigma}$
p-RAI,wr	$\dfrac{\mu_{S_p}\left(1-L(\sigma)\right)}{L(\sigma)\left(1-\sigma\right)}$
dp-RAI	$\dfrac{\mu_{S_p}\,L'\left(\frac{1}{\sigma}\right)}{\sigma}$
dp-RAI,wr	$\dfrac{\mu_{S_p}\,\mu_L}{L(\sigma)}$

Table 2.2: Mean effective service times for the partial operation modes.

is no longer the case for the delayed variants as can be seen from the graphs: efficiency of d-RAI operation and deterministic service times is larger than the efficiency for (shifted) Poisson distributed customer service times and d-RAI,wr operation for sufficiently large values of the mean available period.

Figures 2.24 and 2.25 depict the efficiency of the p-RAI, dp-RAI, p-RAI,wr and dp-RAI,wr operation modes as a function of the relative mean part size $\mu_L/\mu_S = 1/\mu_{S_p}$. The parts share a common shifted symmetrical binomial distribution (figure 2.24) or a shifted geometric distribution (figure 2.25). The distribution of the number of parts per customer does not influence the efficiency given its mean value. The mean customer service time equals 10 slots for both figures. As expected, an increase of the relative mean part size – that is, the parts become larger – implies that the efficiency decreases as longer parts may need to repeat service. Again, the efficiency of the operation modes with resampling is better than the corresponding modes without resampling. Also, as expected, non-delayed modes outperform the equivalent delayed modes. Further, one may note that for geometrically distributed part lengths, the efficiency for p-RAI,wr equals 100% independent of the relative part size. This again follows from the memoryless property of the geometrical distribution. That is, the probability distribution of the remaining service time of an interrupted part equals the distribution of a new sample.

mode	variance of the effective service times
p-RAI,wr	$\frac{{\sigma_{S_p}}^2\,(1-L(\sigma))^2+\mu_{S_p}\left((1-L(\sigma))\,(1+L(\sigma)\,\sigma)-2\,L'(\sigma)\,\sigma(1-\sigma)\right)}{(1-\sigma)^2\,L(\sigma)^2}$
p-RAI	$\frac{\left\{\begin{array}{c}{\sigma_{S_p}}^2\left(1-L\left(\frac{1}{\sigma}\right)\right)^2\sigma+\mu_{S_p}\left(2\,L\left(\frac{1}{\sigma^2}\right)-L\left(\frac{1}{\sigma}\right)^2\right)\sigma\\ -\mu_{S_p}(1-\sigma)\left(L\left(\frac{1}{\sigma}\right)\sigma+2\,L'\left(\frac{1}{\sigma}\right)\right)-\mu_{S_p}\,\sigma^2\end{array}\right\}}{(1-\sigma)^2\,\sigma}$
dp-RAI,wr	$\frac{{\sigma_{S_p}}^2\,{\mu_L}^2+\mu_{S_p}\left(L(\sigma)\left({\sigma_L}^2+{\mu_L}^2\right)+{\mu_L}^2-2\,\mu_L\,\sigma\,L'(\sigma)\right)}{L(\sigma)^2}$
dp-RAI	$\frac{\left\{\begin{array}{c}\mu_{S_p}\left(2\,L'\left(\frac{1}{\sigma^2}\right)-L'\left(\frac{1}{\sigma}\right)^2-L''\left(\frac{1}{\sigma}\right)\right)\sigma^2\\ +{\sigma_{S_p}}^2\,L'\left(\frac{1}{\sigma}\right)^2\sigma^2+\mu_{S_p}\left(2\,L''\left(\frac{1}{\sigma^2}\right)-L'\left(\frac{1}{\sigma}\right)\sigma^3\right)\end{array}\right\}}{\sigma^4}$

Table 2.3: Variance of the effective service times for the partial operation modes.

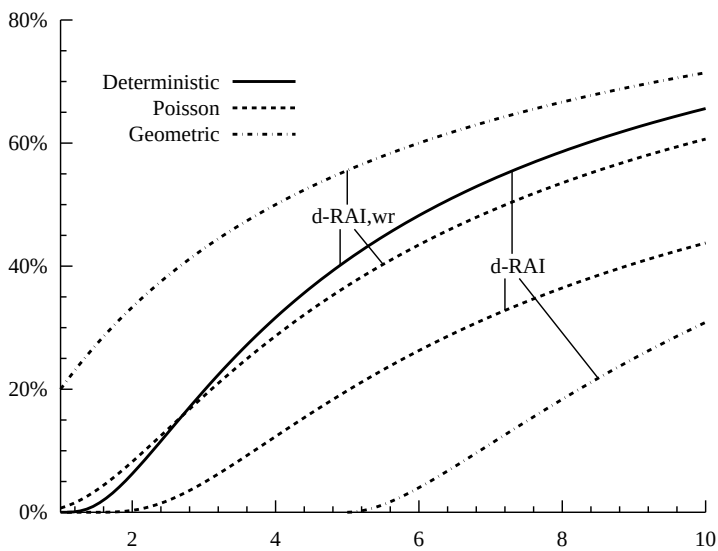


Figure 2.22: Efficiency ϵ vs. the mean A-period μ_A for deterministically, Poisson en geometrically distributed customer service times. (The mean customer service time μ_S equals 5 slots.)

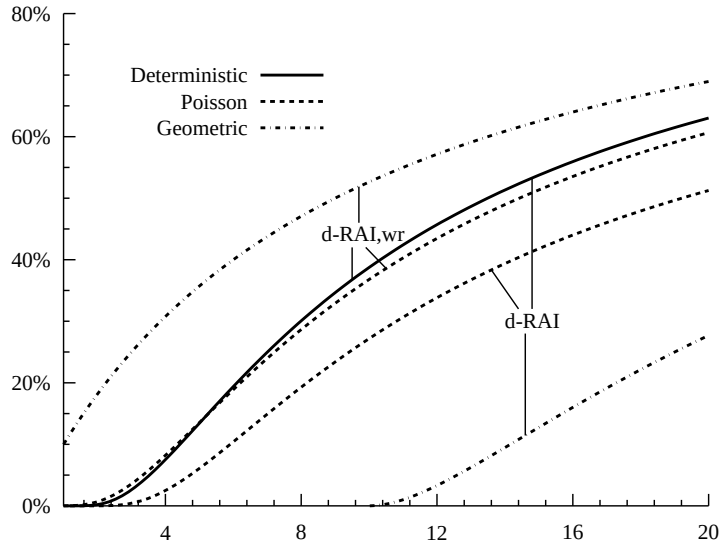


Figure 2.23: Efficiency ϵ vs. the mean A-period μ_A for deterministically, Poisson en geometrically distributed customer service times. (The mean customer service time μ_S equals 10 slots.)

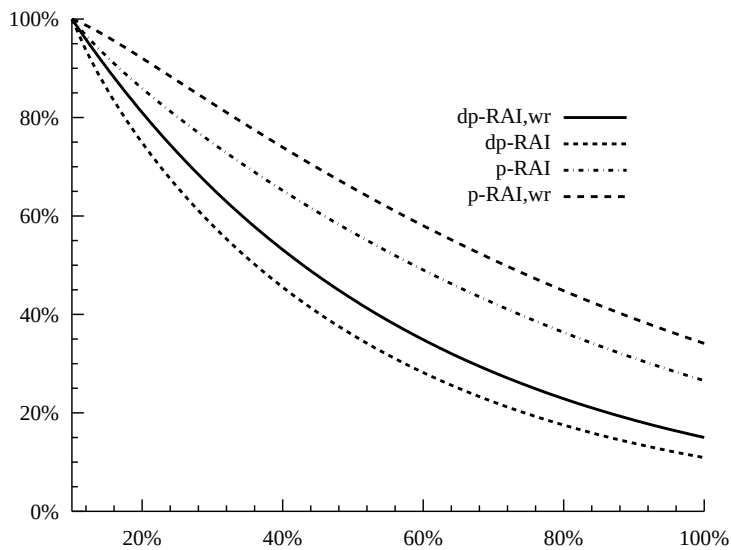


Figure 2.24: Efficiency ϵ vs. the relative part size. (binomially distributed parts, mean customer service time of 10 slots.)

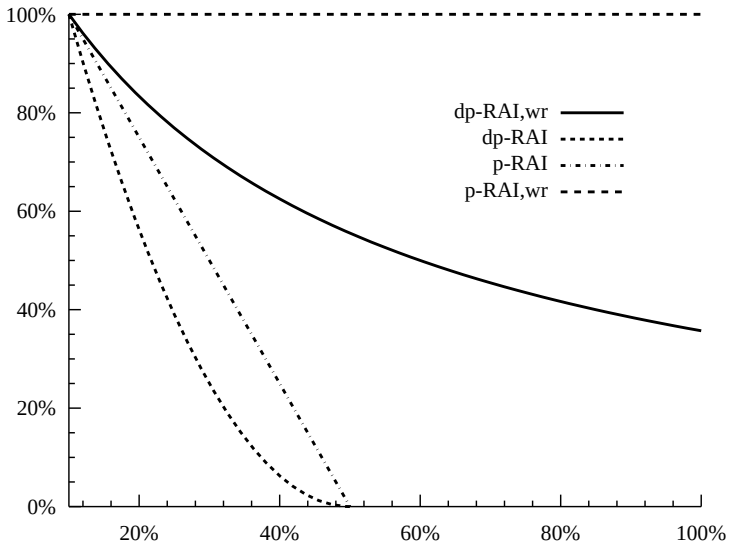


Figure 2.25: Efficiency ϵ vs. the relative part size. (shifted geometrically distributed parts, mean customer service time of 10 slots.)

Chapter 3

Other vacation queues

The queueing models of the former chapter shared the property that the server takes vacations independently of the state of the system. We now consider queueing models with vacation processes that depend in some way on the state of the system. That is, leaving for a vacation at some point in time may depend on the number of customers in the queue at that time, the remaining service time of the customer in service, etc.

In literature, one traditionally distinguishes the following classes of vacation systems: For *exhaustive systems*, the server only takes a vacation whenever there are no customers in the system. A *gated system* keeps on serving until there are no customers left that arrived before the end of the last vacation. *Number-limited systems* pose a maximum number of customers that can be served between server vacations, whereas *time-limited systems* pose a maximum time between vacations, either in a preemptive or a non-preemptive way. For both number- and time-limited systems, the server also takes a vacation whenever there are no more customers to serve before the respective maxima are reached. Additionally, one distinguishes multiple vacation systems and single vacation systems. In a multiple vacation system, the server takes another vacation if there are no customers in the system when it returns from a vacation. In the single vacation system, the server waits for the first arrival when it returns from a vacation. These vacation models are extensively analysed by Takagi [1991, 1993], both in a continuous-time and in a discrete-time setting. We here survey some recent extensions.

Exhaustive vacation systems

Extending the classical exhaustive multiple vacation queue, Li and Zhu [1996] consider an exhaustive vacation system in which the server does not leave immediately for a vacation when the queue becomes empty. In their model, the server first waits for

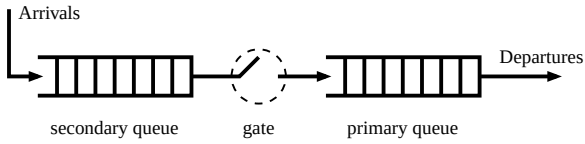


Figure 3.1: Gated vacation system.

a random amount of time and only leaves for a vacation if no customers arrive during this time. Further, Takagi [1994] and Frey and Takahashi [1997] consider exhaustive multiple vacation queueing models with finite population and finite queue capacity respectively. Artalejo [1997] on the other hand investigates a retrial queueing system with exhaustive vacations. In this case, the server leaves for a vacation whenever there are no more customers in "orbit". That is, the server leaves when there are no more customers that "retry" to receive service from time to time. All these queueing systems are investigated in a continuous-time setting. The arrival processes are Poisson processes and the service times constitute series of i.i.d. random variables. Related is also Servi's continuous-time system [2002] in which the server does not leave for a vacation but temporarily slows down when the system becomes empty. The arrival process is again a Poisson process and service times share an exponential distribution.

Gated vacation systems

A number of different authors investigate gated vacation queues with feedback – i.e., a fraction of the customers reenter the queue after service – in a continuous-time setting. Takagi [1991] already considered the gated multiple vacation queueing system with Bernoulli feedback in his monograph. A system with a more complex feedback process (with multiple types of feedback) and with a Poisson arrival process is analysed by Boxma and Yechiali [1997] as well as by Choi et al. [2003]. The former authors consider the longest present (customers that are fed back get priority) service discipline while the latter authors consider the FIFO scheduling discipline. Lee [1997b] considers a queueing system with a more general arrival process, with Bernoulli feedback and a randomly gated vacation discipline. Here, the server may leave for another vacation while there are customers waiting that arrived before the end of the last vacation.

One may represent a gated vacation system as one with two queues separated by a gate as depicted in figure 3.1. Customers arrive in the secondary queue and wait in front of the gate. The latter opens when the server returns from a vacation and customers move in batch to the primary queue. Customers in this queue then receive service according to some service discipline. Some authors also consider this type of system without vacations. Ishizaki et al. [1994] considers a discrete-time queueing system with a gate where the gate opening intervals constitute a series of i.i.d. random

variables. The same authors [1995] also consider this system in the case that arrivals occur in both the secondary and the primary queue. In this case, there is a kind of service differentiation (priority queueing) between arrivals in the primary queue and arrivals in the secondary queue.

Time-limited systems

Time-limited systems may either operate gated or non-gated. For a gated time-limited system, the server takes a vacation when the timer expires or when there are no more customers in the system that arrived before the last vacation. The non-gated vacation system takes a vacation when the timer expires or when there are no more customers in the system. Further, when the timer expires, the server either leaves immediately for a vacation (a preemptive schedule) or leaves for a vacation after completing the service of the customer in service (a non-preemptive schedule). In the case the server leaves immediately, the customer in service either resumes his service (preemptive resume), repeats his service (preemptive repeat) or repeats his service with a different service time sample (preemptive repeat with resampling) after the vacation. Notice that the same types of scheduling disciplines are also used in the context of priority queueing systems.

Amongst others, Leung and Eisenberg [1991], Leung and Lucantoni [1994] and Alfa [1995] consider continuous-time non-gated time-limited vacation queueing systems. Both Leung and Eisenberg [1991] and Leung and Lucantoni [1994] assume a Poisson arrival process and generally distributed service times and vacation times. The former authors consider a preemptive resume time-limited vacation system whereas the latter consider the non-preemptive case. Further, Leung and Lucantoni [1994] also allow correlation between vacations and timers. Alfa [1995] on the other hand, studies a non-gated preemptive resume time-limited vacation system with a Markovian arrival process and phase-type service times and vacations.

A gated continuous-time time-limited vacation system is investigated by Leung and Eisenberg [1990]. These authors consider a preemptive resume schedule. The arrival process is a Poisson process and service times and vacation periods are generally distributed. Further, Shomrony and Yechiali [2001] investigate the continuous-time non-preemptive and preemptive repeat with resampling time-limited gated vacation queueing systems with a Poisson batch arrival process. Related are also the randomly timed gated polling models considered by Eliazar and Yechiali [1998b,a]. These authors investigate gated polling policies corresponding to the preemptive resume, the preemptive repeat with resampling and the non-preemptive cases described above, again in a continuous-time setting.

The rest of this chapter is organised as follows. In the following section we consider the so-called *gated-exhaustive* vacation queue which incorporates properties of both the gated and the exhaustive vacation system. In section 3.2, we present a Markovian

vacation model which encapsulates numerous classical non-gated vacation systems. In particular, this model allows that the server interrupts a customer's service to leave for a vacation. As in chapter 2, we again focus on the continue after interruption, the repeat after interruption and the repeat after interruption with resampling modes.

3.1 The gated-exhaustive vacation queue

This section deals with a queueing system with vacations that encapsulates both the gated and the exhaustive vacation systems. Recall that a gated system can be represented by the two queue system depicted in figure 3.1. In the classical gated system, arrivals are only allowed in the secondary queue. Here we extend the latter by also allowing direct arrivals in the primary queue. As we will see further, our model encapsulates the exhaustive and gated vacation systems with single and multiple vacations.

As in section 2.2, we here apply the method of the supplementary variable. The rest of this section follows the lines of our contribution [Fiems et al., 2002a].

3.1.1 Mathematical model

We first give a more detailed description of the queueing model under consideration.

We again investigate a discrete-time queueing system. There is a single server and two infinite capacity queues separated by a gate as depicted in figure 3.2. During the consecutive slots, customers arrive either in the primary or in the secondary queues. Customers in the primary queue are served according to a FIFO queueing discipline whereas arrivals in the secondary queue move in order of arrival to the primary queue whenever the gate opens (see further).

The numbers of customers arriving in the primary and secondary queues during the consecutive slots are modelled by means of two series of from slot to slot independent but jointly dependent random variables with common joint probability mass function $e(m,n)$ $(m,n \geq 0)$ and corresponding joint probability generating function $E(z_1,z_2)$:

$$E(z_1,z_2) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} e(m,n){z_1}^m{z_2}^n. \tag{3.1}$$

For ease of notation, we also introduce the probability generating function $E(z) = E(z,z)$ of the total number of arrivals per slot and the marginal probability generating functions $E_1(z) = E(z,1)$ and $E_2(z) = E(1,z)$ corresponding to the number of arrivals per slot in the primary and secondary queues respectively.

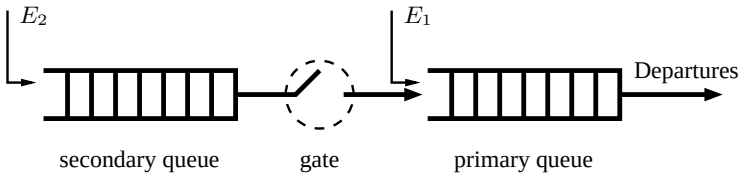


Figure 3.2: The gated-exhaustive vacation system.

As before, the service of customers is synchronised with respect to slot boundaries which implies that service of a customer cannot start before the slot following his arrival slot. The service times of the consecutive customers constitute a series of i.i.d. positive random variables with common probability generating function $S(z)$.

Whenever there are no more primary customers in the system at the beginning of a slot, the server takes a vacation. Upon returning from this vacation the gate opens and all customers in the secondary queue move in batch to the primary queue. If there are no customers in the system upon returning from the vacation, the server immediately takes a new vacation. The lengths (in slots) of the consecutive server vacations are modelled as a series of independent random variables with common probability generating functions $B_1(z)$ or $B_2(z)$ depending on whether the vacation is not immediately or is immediately preceded by another vacation.

Notice that the model under consideration relates to priority queueing systems as customers arriving in the primary queue receive priority over customers arriving in the secondary queue. Only if there are no more customers in the primary queue, waiting (low-priority) customers in the secondary queue can jump to the (high-priority) primary queue and be served.

3.1.2 System equations

Let $V_{r,1}^{(k)}$ and $V_{r,2}^{(k)}$ denote the number of customers in the primary and secondary queues at the beginning of the k-th slot respectively. Further, let $F^{(k)}$ denote the remaining number of vacation slots following slot k if slot k is a vacation slot and let $H^{(k)}$ denote the number of remaining service slots of the customer in service following slot k if this is not the case. As in chapter 2, we let $Q^{(k)}$ denote the number of available servers during slot k. That is, $Q^{(k)}$ equals 0 during vacations and 1 when a customer is served. The system under consideration then yields the following set of system equations.

- If $Q^{(k)} = 0 \wedge F^{(k)} > 0$, then:

$$\begin{aligned} Q^{(k+1)} &= 0, \\ F^{(k+1)} &= F^{(k)} - 1, \\ V_{r,1}^{(k+1)} &= V_{r,1}^{(k)} + E_1^{(k)}, \\ V_{r,2}^{(k+1)} &= V_{r,2}^{(k)} + E_2^{(k)}; \end{aligned} \tag{3.2}$$

  that is, if the server is on vacation during slot k and does not end its vacation at the end of this slot, the server continues its vacation during slot $k+1$.

- If $Q^{(k)} = F^{(k)} = 0 \wedge V_{r,1}^{(k)} + V_{r,2}^{(k)} + E_1^{(k)} + E_2^{(k)} = 0$, then:

$$\begin{aligned} Q^{(k+1)} &= 0, \\ F^{(k+1)} &= B_2 - 1, \\ V_{r,1}^{(k+1)} &= 0, \\ V_{r,2}^{(k+1)} &= 0; \end{aligned} \tag{3.3}$$

  that is, slot k is the last slot of a vacation after which the system is found to be empty. As such, the server immediately takes a new vacation.

- If $Q^{(k)} = F^{(k)} = 0 \wedge V_{r,1}^{(k)} + V_{r,2}^{(k)} + E_1^{(k)} + E_2^{(k)} > 0$, then:

$$\begin{aligned} Q^{(k+1)} &= 1, \\ H^{(k+1)} &= S - 1, \\ V_{r,1}^{(k+1)} &= V_{r,1}^{(k)} + V_{r,2}^{(k)} + E_1^{(k)} + E_2^{(k)}, \\ V_{r,2}^{(k+1)} &= 0; \end{aligned} \tag{3.4}$$

  that is, the server returns from a vacation and finds the system non-empty. Therefore, the gate opens at the end of slot k and a new customer starts service at the beginning of slot $k+1$.

- If $Q^{(k)} = 1 \wedge H^{(k)} > 0$, then:

$$\begin{aligned} Q^{(k+1)} &= 1, \\ H^{(k+1)} &= H^{(k)} - 1, \\ V_{r,1}^{(k+1)} &= V_{r,1}^{(k)} + E_1^{(k)}, \\ V_{r,2}^{(k+1)} &= V_{r,2}^{(k)} + E_2^{(k)}; \end{aligned} \tag{3.5}$$

  that is, there is a customer in service during slot k and the latter continues service during slot $(k+1)$.

- If $Q^{(k)} = 1 \wedge H^{(k)} = 0 \wedge V_{r,1}^{(k)} + E_1^{(k)} \geq 1$, then:

$$\begin{aligned} Q^{(k+1)} &= 1, \\ H^{(k+1)} &= S - 1, \\ V_{r,1}^{(k+1)} &= V_{r,1}^{(k)} + E_1^{(k)} - 1, \\ V_{r,2}^{(k+1)} &= V_{r,2}^{(k)} + E_2^{(k)}; \end{aligned} \tag{3.6}$$

  that is, the customer in service leaves the system at the end of slot k and – as there are other customers in the primary queue – a new customer starts service at the beginning of slot $k+1$.

- If $Q^{(k)} = V_{r,1}^{(k)} = 1 \wedge H^{(k)} = E_1^{(k)} = 0$, then:

$$\begin{aligned} Q^{(k+1)} &= 0, \\ F^{(k+1)} &= B_1 - 1, \\ V_{r,1}^{(k+1)} &= 0, \\ V_{r,2}^{(k+1)} &= V_{r,2}^{(k)} + E_2^{(k)}; \end{aligned} \tag{3.7}$$

  that is, the customer in service leaves the system at the end of slot k and there are no other customers in the primary queue. Therefore, the server stops serving customers and leaves for a vacation.

In the former set of system equations, $E_1^{(k)}$ and $E_2^{(k)}$ denote the number of arrivals during slot k in the primary and secondary queues respectively whereas S denotes the service time of a random customer. Further, B_1 and B_2 denote the lengths of a random server vacation not immediately or immediately preceded by another vacation respectively. The corresponding probability generating functions are denoted by $E_1(z)$, $E_2(z)$, $S(z)$, $B_1(z)$ and $B_2(z)$.

It can be observed from the system equations (3.2) to (3.7) above that the series of vectors $\{V_{r,1}^{(k)}, V_{r,2}^{(k)}, F^{(k)}$ or $H^{(k)}, Q^{(k)}\}$ forms a four-dimensional Markov chain. The variable $F^{(k)}$ only has meaning if $Q^{(k)} = 0$ and $H^{(k)}$ only if $Q^{(k)} = 1$. As such, the kth vector provides a complete description of the state of the system at a slot k. In what follows, we show how to obtain the steady-state distribution of the system state, both during busy slots (slots during which a customer receives service) and during vacation slots.

3.1.3 The joint probability generating functions

Let $P_1^{(k)}(x, z_1, z_2)$ and $P_2^{(k)}(x, z_1, z_2)$ denote the following (partial) joint probability generating functions:

$$\begin{aligned} P_1^{(k)}(x, z_1, z_2) &= \mathrm{E}\left[x^{H^{(k)}} z_1^{V_{r,1}^{(k)}-1} z_2^{V_{r,2}^{(k)}} \middle| Q^{(k)} = 1 \right] \Pr\left[Q^{(k)} = 1 \right], \\ P_2^{(k)}(x, z_1, z_2) &= \mathrm{E}\left[x^{F^{(k)}} z_1^{V_{r,1}^{(k)}} z_2^{V_{r,2}^{(k)}} \middle| Q^{(k)} = 0 \right] \Pr\left[Q^{(k)} = 0 \right]. \end{aligned} \tag{3.8}$$

I.e., the former partial probability generating functions correspond to the system state during busy slots and vacation slots respectively.

The preceding system equations (3.2) to (3.7) can be used to relate these generating functions for slot $(k+1)$ to those for slot k. After some standard z-transform manipulations, we find

$$\begin{aligned} P_1^{(k+1)}(x, z_1, z_2) = &\frac{S(x)}{x\, z_1} \left(P_2^{(k)}(0, z_1, z_1) E(z_1) - P_2^{(k)}(0,0,0) E(0) \right) \\ &+ \frac{S(x)}{x\, z_1} \left(P_1^{(k)}(0, z_1, z_2)\, E(z_1, z_2) - P_1^{(k)}(0, 0, z_2)\, E(0, z_2) \right) \\ &+ \frac{E(z_1, z_2)}{x} \left(P_1^{(k)}(x, z_1, z_2) - P_1^{(k)}(0, z_1, z_2) \right) \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} P_2^{(k+1)}(x, z_1, z_2) = &\frac{E(z_1, z_2)}{x} \left(P_2^{(k)}(x, z_1, z_2) - P_2^{(k)}(0, z_1, z_2) \right) \\ &+ \frac{B_2(x)}{x}\, E(0)\, P_2^{(k)}(0,0,0) \\ &+ \frac{B_1(x)}{x}\, E(0, z_2)\, P_1^{(k)}(0, 0, z_2). \end{aligned} \tag{3.10}$$

It can be shown that the gated-exhaustive vacation system under consideration reaches steady state when the offered load ρ is less than 1, that is, when

$$\rho = (\mu_{E_1} + \mu_{E_2})\, \mu_S < 1. \tag{3.11}$$

Here μ_{E_1} and μ_{E_2} denote the mean number of arrivals in a slot in the primary and secondary queue respectively. As before, μ_S denotes the mean service time.

Under the assumption that the system reaches steady state, let $P_i(x, z_1, z_2)$ $(i = 1, 2)$ denote the steady-state probability generating functions, i.e., $P_i(x, z_1, z_2) =$

$\lim_{k\to\infty} P_i^{(k)}(x, z_1, z_2)$ for $i = 1, 2$. Letting $k \to \infty$ in (3.9) and (3.10), and solving for $P_1(x, z_1, z_2)$ and $P_2(x, z_1, z_2)$ respectively then yields

$$P_1(x, z_1, z_2) = \frac{\left\{\begin{array}{l} S(x)P_2(0, z_1, z_1)E(z_1) - S(x)E(0, z_2)P_1(0, 0, z_2) \\ - E(z_1, z_2)(z_1 - S(x))P_1(0, z_1, z_2) \\ - S(x)P_2(0, 0, 0)E(0) \end{array}\right\}}{(x - E(z_1, z_2))\, z_1}, \tag{3.12}$$

and

$$P_2(x, z_1, z_2) = \frac{\left\{\begin{array}{l} E(0)B_2(x)P_2(0, 0, 0) - E(z_1, z_2)P_2(0, z_1, z_2) \\ + E(0, z_2)B_1(x)P_1(0, 0, z_2) \end{array}\right\}}{x - E(z_1, z_2)}. \tag{3.13}$$

Both partial probability generating functions are bounded for $|x|, |z_1|, |z_2| \leq 1$. Therefore, the numerator on the right hand side of both equations (3.12) and (3.13) vanishes for $(E(z_1, z_2), z_1, z_2)$ $(|z_1|, |z_2| \leq 1)$ as the corresponding denominators vanish as well. This observation yields:

$$P_1(0, z_1, z_2) = S(E(z_1, z_2)) \frac{\left\{\begin{array}{l} E(z_1)P_2(0, z_1, z_1) - E(0)P_2(0, 0, 0) \\ - E(0, z_2)P_1(0, 0, z_2) \end{array}\right\}}{E(z_1, z_2)\,(z_1 - S(E(z_1, z_2)))} \tag{3.14}$$

and

$$\begin{aligned} P_2(0, z_1, z_2) = {} & \frac{E(0)B_2(E(z_1, z_2))}{E(z_1, z_2)} P_2(0, 0, 0) \\ & + \frac{E(0, z_2)B_1(E(z_1, z_2))}{E(z_1, z_2)} P_1(0, 0, z_2). \end{aligned} \tag{3.15}$$

Equations (3.12) to (3.15) then yield the following expressions for the (partial) probability generating functions,

$$P_1(x, z_1, z_2) = \frac{S(x) - S(E(z_1, z_2))}{z_1 - S(E(z_1, z_2))} \frac{\left\{\begin{array}{l} B_1(E(z_1))E(0, z_1)P_1(0, 0, z_1) \\ - E(0)(1 - B_2(E(z_1)))P_2(0, 0, 0) \\ - P_1(0, 0, z_2)E(0, z_2) \end{array}\right\}}{x - E(z_1, z_2)} \tag{3.16}$$

and

$$P_2(x,z_1,z_2) = \frac{\left\{\begin{array}{l} E(0,z_2)(B_1(x) - B_1(E(z_1,z_2)))P_1(0,0,z_2) \\ + E(0)(B_2(x) - B_2(E(z_1,z_2)))P_2(0,0,0) \end{array}\right\}}{x - E(z_1,z_2)}. \tag{3.17}$$

That is, the generating functions $P_1(x,z_1,z_2)$ and $P_2(x,z_1,z_2)$ are completely determined once one has determined $P_1(0,0,z)$ and $P_2(0,0,0)$.

The quantity $P_2(0,0,0)$ can directly be obtained from the normalisation condition $P_1(1,1,1) + P_2(1,1,1) = 1$. Using de l'Hôpital's rule, this yields

$$P_2(0,0,0) = \frac{1 - \rho - \mu_{B_1} E_1(0) P_1(0,0,1)}{E(0)\mu_{B_2}}. \tag{3.18}$$

Here, μ_{B_1} and μ_{B_2} denote the mean vacation length given that the vacation is not or is preceded immediately by a vacation respectively. One may also obtain the former expression directly as follows. The fraction of busy slots is given by the load ρ. As the system is either busy or on vacation, the fraction of vacation slots is therefore given by $1 - \rho$. On the other hand, $P_2(0,0,0)E(0)$ denotes the probability that a new vacation starts at the end of a random slot and that this vacation is immediately preceded by another vacation whereas $P_1(0,0,1)E_1(0)$ denotes the probability that a new vacation starts at the end of a random slot and that this vacation is not immediately preceded by another vacation. The fraction of vacation slots is therefore also given by $P_1(0,0,1)E_1(0)\mu_{B_1} + P_2(0,0,0)E(0)\mu_{B_2}$. Comparison of these fractions then yields the former expression for $P_2(0,0,0)$.

Finally, to characterise the last unknown function $P_1(0,0,z)$, we can proceed as follows. First notice that for $\rho < 1$, Rouché's theorem assures for each $|z_2| \leq 1$ the existence of a unique value $\chi(z_2)$ such that, $|\chi(z_2)| \leq 1$ and

$$\chi(z_2) = S(E(\chi(z_2), z_2)). \tag{3.19}$$

The partial probability generating function $P_1(0,z_1,z_2)$ is bounded for $|z_1|, |z_2| \leq 1$. The numerator on the right hand side of (3.14) therefore vanishes for all values $(0, \chi(z_2), z_2)$, $|z_2| \leq 1$ as the corresponding denominator vanishes for these values. This observation and equation (3.15) then yield:

$$\begin{aligned} P_1(0,0,z)E(0,z) = {} & P_1(0,0,\chi(z))E(0,\chi(z))B_1(E(\chi(z))) \\ & + P_2(0,0,0)E(0)\left[B_2(E(\chi(z))) - 1\right]. \end{aligned} \tag{3.20}$$

To simplify notation, let $\Omega(z)$ denote,

$$\Omega(z) \triangleq \frac{E(0,z)\,P_1(0,0,z)}{E_1(0)\,P_1(0,0,1)}. \tag{3.21}$$

One sees that by definition $P_1(0,0,z)/P_1(0,0,1)$ is the probability generating function of the secondary queue content at the beginning of slots where there is only one customer in the primary queue and where this customer leaves at the end of this slot. Such a slot is the last slot of a contiguous period of busy slots (a busy period) whenever there are no customer arrivals in the primary queue during this slot. Therefore, one verifies easily that $\Omega(z)$ is the probability generating function of the number of customers in the system at the end of a busy period. This follows from the fact that $E(0,z)/E_1(0)$ denotes the probability generating function of the number of customer arrivals in the secondary queue given that there are no arrivals in the primary queue.

Combining equations (3.20) and (3.21), we get,

$$\Omega(\chi(z)) = \frac{\Omega(z) + \kappa\,[1 - B_2(E(\chi(z)))]}{B_1(E(\chi(z)))}, \tag{3.22}$$

with

$$\kappa \triangleq \frac{E(0)\,P_2(0,0,0)}{E_1(0)\,P_1(0,0,1)} = \Omega(0)\,\frac{B_1(E(0))}{1 - B_2(E(0))}. \tag{3.23}$$

The second equality in the former equation follows from equations (3.15) and (3.21).

Now, consider the series, $x_0 = 0$, $x_i = \chi(x_{i-1})$ for $i > 0$. This series clearly converges to 1 as long as $\chi'(1) < 1$ which is implied by the steady-state condition $\rho < 1$. The convergence is brought about by the fact that $\chi(z)$ proves to be a probability generating function. In particular, it is the probability generating function of the number of secondary customers that arrive during the sub-busy period of a primary customer (see e.g., Walraevens et al. [2002]). The *sub-busy period* of a primary customer is defined to start from the moment this customer enters service, and ends on the first occasion in which the (primary) queue contains one customer less than at the beginning of the sub-busy period.

As the series $\{x_i\}$ converges to 1, one observes that the series $y_i \triangleq \kappa/\Omega(x_i)$, $i \geq 0$ converges to κ. This follows from the normalisation condition of the probability generating function $\Omega(z)$. From equation (3.23), one easily obtains

$$y_0 \triangleq \frac{\kappa}{\Omega(0)} = \frac{B_1(E(0))}{1 - B_2(E(0))}, \tag{3.24}$$

whereas equation (3.22) yields

$$y_{i+1} = \frac{B_1(E(x_{i+1}))\, y_i}{1 + (1 - B_2(E(x_{i+1})))\, y_i}. \tag{3.25}$$

This recursive relation then allows us to calculate κ numerically up to any desired precision since, as mentioned, $y_i \to \kappa$ as $i \to \infty$.

Summarising, we get

$$P_1(x, z_1, z_2) = \frac{1-\rho}{\mu_{B_1} + \kappa\mu_{B_2}} \frac{S(x) - S(E(z_1, z_2))}{z_1 - S(E(z_1, z_2))} \frac{\left\{\begin{array}{l} B_1(E(z_1))\Omega(z_1) - \Omega(z_2) \\ -\kappa(1 - B_2(E(z_1))) \end{array}\right\}}{x - E(z_1, z_2)} \tag{3.26}$$

and

$$P_2(x, z_1, z_2) = \frac{1-\rho}{\mu_{B_1} + \kappa\mu_{B_2}} \frac{\left\{\begin{array}{l} \Omega(z_2)(B_1(x) - B_1(E(z_1, z_2))) \\ +\kappa(B_2(x) - B_2(E(z_1, z_2))) \end{array}\right\}}{x - E(z_1, z_2)}, \tag{3.27}$$

where $\Omega(z)$ is implicitly defined by (3.22) and where one can determine κ numerically using equations (3.24) and (3.25).

3.1.4 Queue content at various epochs

We now show how the partial probability generating functions $P_i(x, z_1, z_2)$ $(i = 1, 2)$ allow us to evaluate the performance of the system. Let $V_r(z_1, z_2)$ denote the joint probability generating function of the numbers of customers in the primary and secondary queues at random slot boundaries, then, using the definitions (3.8), we get

$$V_r(z_1, z_2) = z_1\, P_1(1, z_1, z_2) + P_2(1, z_1, z_2). \tag{3.28}$$

Plugging in equations (3.26) and (3.27) in the former expression, we get

$$V_r(z_1, z_2) = (1-\rho) \frac{\left\{\begin{array}{l} z_1(1 - S(E(z_1, z_2)))(B_2(E(z_1))\kappa + B_1(E(z_1))\Omega(z_1)) \\ -(1 - z_1)S(E(z_1, z_2))(\Omega(z_2) + \kappa) \\ -z_1(B_1(E(z_1, z_2))\Omega(z_2) + B_2(E(z_1, z_2))\kappa) \\ +S(E(z_1, z_2))(B_1(E(z_1, z_2))\Omega(z_2) + B_2(E(z_1, z_2))\kappa) \end{array}\right\}}{(\mu_{B_1} + \kappa\mu_{B_2})(z_1 - S(E(z_1, z_2)))(1 - E(z_1, z_2))}. \tag{3.29}$$

There is a customer departure at the end of a random busy slot when the customer in service does not require more service slots after the present slot. The departing customer then leaves behind all customers in the system at the beginning of this random busy slot as well as all customer arrivals during this slot. Let $V_{r,1}$, $V_{r,2}$ denote the primary and the secondary queue content immediately after a random slot boundary respectively and let Q and H denote the number of available servers during a random slot and the remaining service time of the customer in service following a random busy slot respectively. The joint probability generating function of the primary and the secondary queue content at departure epochs is then given by

$$V_d(z_1, z_2) = \mathrm{E}\left[z_1^{V_{r,1}+E_1-1} z_2^{V_{r,2}+E_2} | H = 0 \wedge Q = 1\right]. \tag{3.30}$$

In view of the definitions (3.8) and the independent nature of the arrival process, we find the following expression for the joint probability generating function of the numbers of customers in the primary and secondary queues at departure epochs:

$$V_d(z_1, z_2) = \frac{P_1(0, z_1, z_2)}{P_1(0, 1, 1)} E(z_1, z_2). \tag{3.31}$$

Plugging in expression (3.26) then yields:

$$\begin{aligned} &V_d(z_1, z_2) = \\ &\quad \frac{(1-\rho)S(E(z_1, z_2))}{\mu_{B_1} + \kappa\mu_{B_2}} \frac{B_1(E(z_1))\Omega(z_1) - \Omega(z_2) - \kappa + B_2(E(z_1))\kappa}{(z_1 - S(E(z_1, z_2)))\,\mu_E}. \end{aligned} \tag{3.32}$$

Using the moment generating property of probability generating functions, one easily obtains performance measures such as the mean and the variance of the numbers of customers in the primary and secondary queues and the correlation factor between the numbers of customers in the primary and secondary queues.

From (3.29), one may in particular retrieve the probability generating function of the total queue content at random slot boundaries $V_r(z) \triangleq V_r(z, z)$,

$$V_r(z) = \frac{(1-\rho)S(E(z))(z-1)}{z - S(E(z))} \frac{(1 - B_1(E(z)))\Omega(z) + \kappa(1 - B_2(E(z))}{(1 - E(z))\,(\mu_{B_1} + \kappa\,\mu_{B_2})}. \tag{3.33}$$

Close observation of the former expression reveals that the probability generating function of the total queue content is the product of the probability generating function of the queue content at random slot boundaries of a queueing system without vacations (see section 1.2) and of the generating function of the total queue content at the beginning of a random vacation slot $V_v(z)$:

$$V_v(z) = \frac{P_2(1, z, z)}{P_2(1, 1, 1)} = \frac{(1 - B_1(E(z)))\Omega(z) + \kappa(1 - B_2(E(z))}{(1 - E(z))\,(\mu_{B_1} + \kappa\,\mu_{B_2})}. \tag{3.34}$$

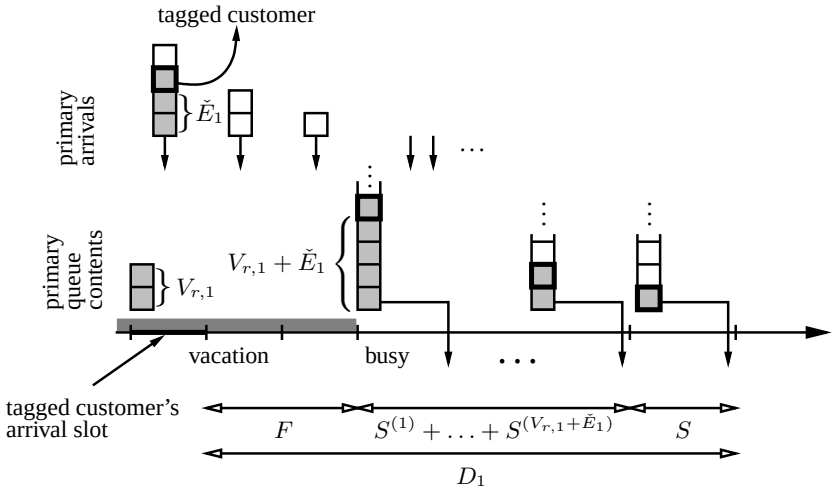


Figure 3.3: Delay experienced by an arbitrary but tagged customer arriving in the primary queue during a vacation slot. Customers that are served during the delay time of the tagged customer are shaded.

This property is known as the stochastic decomposition of vacation queueing systems. A similar result is known to hold for a large class of continuous-time queueing systems (see a.o., Fuhrmann and Cooper [1985], Takagi [1991, 1993]).

3.1.5 Customer delay

Let D_1 and D_2 denote the delay of a random customer that arrives in the primary and secondary queues respectively and let $D_1(z)$ and $D_2(z)$ denote the corresponding probability generating functions. Further, $V_{r,1}$ and $V_{r,2}$ denote the primary and secondary queue content at the beginning of a random customer's arrival slot and H (F) denotes the remaining service (vacation) time after a random customer's arrival slot given that the server is busy (on vacation) during this slot.

Delay of a random customer arriving in the primary queue

We first consider a random "but tagged" customer that arrives in the primary queue. Such a customer receives service when all customers that arrived in the primary queue before the tagged customer are served. Given that the customer under consideration arrives during a vacation slot, his delay equals the remaining vacation time, augmented by the sum of the service times of the customers present in the primary queue upon arrival of the tagged customer and by his own service time as depicted in figure 3.3.

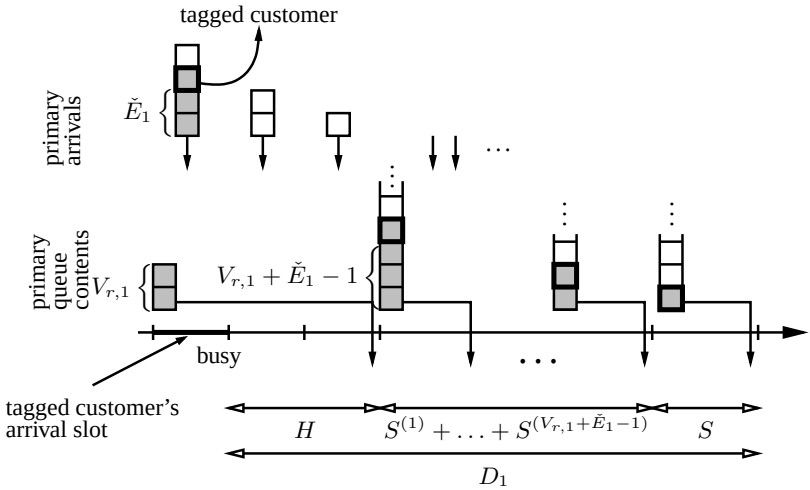


Figure 3.4: Delay experienced by an arbitrary but tagged customer arriving in the primary queue during a busy slot. Customers that are served during the delay time of the tagged customer are shaded.

We get:

$$D_1 = F + S + \sum_{j=1}^{V_{r,1}+\check{E}_1} S^{(j)}. \tag{3.35}$$

Here $\check{E}_1$ denotes the number of arrivals in the primary queue during the tagged customer's arrival slot that are served before this customer. Further, $S^{(j)}$ denotes the service time of the j-th customer starting service after the tagged customer's arrival slot and S denotes the tagged customer's own service time.

Similarly, given that the customer arrives during a busy slot, his delay equals the remaining service time of the customer in service upon arrival of the tagged customer, augmented by the sum of the service times of the customers present in the primary queue upon arrival of the tagged customer (including the customers that arrive at the sale instant but that are served before the tagged customer) and by his own service time as depicted in figure 3.4. We find:

$$D_1 = H + S + \sum_{j=1}^{V_{r,1}+\check{E}_1-1} S^{(j)}. \tag{3.36}$$

Due to the independent nature of the arrival process, the state of the system at the beginning of a random customer's arrival slot, has the same stochastic properties as the

state of the system at a random slot boundary. That is, equations (3.26) and (3.27) also denote the probability generating functions of the state of the system at the beginning of a random customer's arrival slot. One then easily obtains the probability generating function of the delay of customers that arrive in the primary queue;

$$D_1(z) = (P_1(z, S(z), 1) + P_2(z, S(z), 1))\ \check{E}_1(S(z))\, S(z). \tag{3.37}$$

Here $\check{E}_1(z)$ denotes the probability generating function corresponding to $\check{E}_1$ as derived in section 1.2:

$$\check{E}_1(z) = \frac{E_1(z) - 1}{\mu_{E_1}\,(z-1)}. \tag{3.38}$$

Plugging equations (3.26), (3.27) and (3.38) into (3.37), we obtain an expression for the probability generating function $D_1(z)$ in terms of the system parameters:

$$D_1(z) = \frac{(1-\rho)S(z)}{\mu_{B_1} + \kappa\mu_{B_2}} \frac{E_1(S(z)) - 1}{\mu_{E_1}(S(z) - 1)} \cdot \frac{\left\{ \begin{array}{l} \kappa(B_2(z) - B_2(E_1(S(z))) - 1 + B_2(E(S(z)))) \\ + B_1(z) - B_1(E_1(S(z))) + B_1(E(S(z)))\Omega(S(z)) - 1 \end{array} \right\}}{z - E_1(S(z))}. \tag{3.39}$$

The moment generating property of probability generating functions then yields various moments of the delay of customers arriving in the primary queue.

Delay of a random customer arriving in the secondary queue

Finding the delay D_2 of a random (tagged) customer that arrives in the secondary queue is somewhat more involved. Figure 3.5 illustrates the case where this tagged customer arrives during a vacation slot. The delay of the tagged customer includes the remaining slots of the vacation during which this customer arrives. At the end of this vacation, the gate opens and the tagged customer joins the primary queue. Clearly, all customers that are in the primary queue when the gate opens are served before the tagged customer. Also, all customers that are present in the secondary queue upon arrival of the tagged customer are served before the latter. The tagged customer's delay includes the service times of all these customers. These observations then yield the following expression for the tagged customer's delay D_2:

$$D_2 = F + S + \sum_{j=1}^{\check{V}_1} S^{(j)}, \tag{3.40}$$

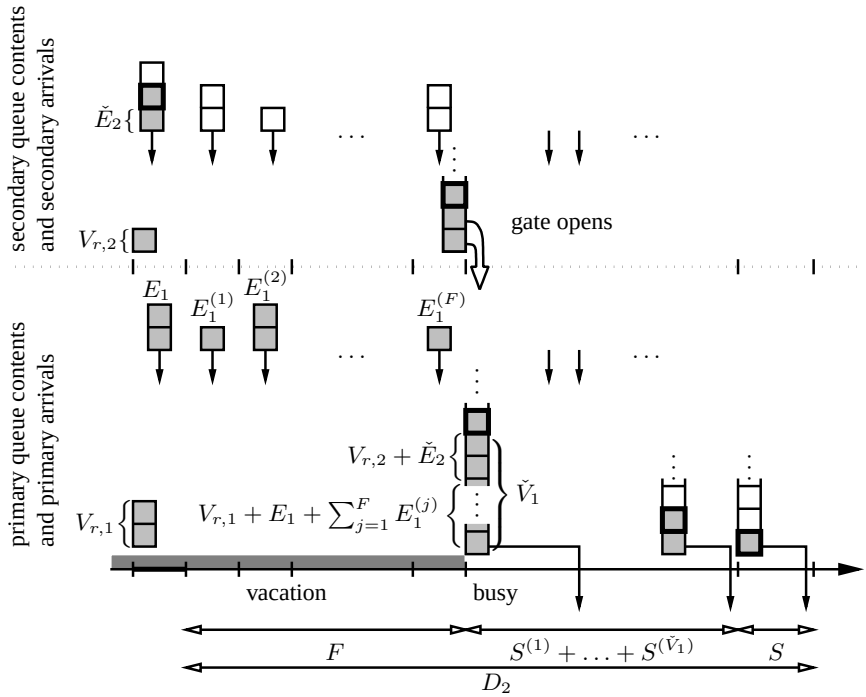


Figure 3.5: Delay experienced by a tagged customer arriving in the secondary queue during a vacation slot.

where

$$\check{V}_1 = V_{r,1} + V_{r,2} + E_1 + \check{E}_2 + \sum_{j=1}^{F} E_1^{(j)} \tag{3.41}$$

denotes the number of customers that are served before the tagged customer. Here, $E_1^{(j)}$ denotes the number of arrivals in the primary queue during the j-th slot following the tagged customer's arrival slot, whereas E_1 and $\check{E}_2$ denote the number of arrivals in the primary queue during the tagged customer's arrival slot and the number of arrivals in the secondary queue during this slot that are served before the tagged customer respectively.

Given that the server is busy during the tagged customer's arrival slot (see figure 3.6), the customer can only be served after the next vacation. As the server only leaves for a vacation whenever the primary queue is empty, we find that the tagged customer's delay includes the number of slots it takes to empty the primary queue. Let $\hat{V}_1$ denote the primary queue content upon departure of the customer in service during the tagged customers arrival slot. The latter equals the primary queue content $V_{r,1}$ at the beginning of the tagged customer's arrival slot, augmented with the arrivals in the primary

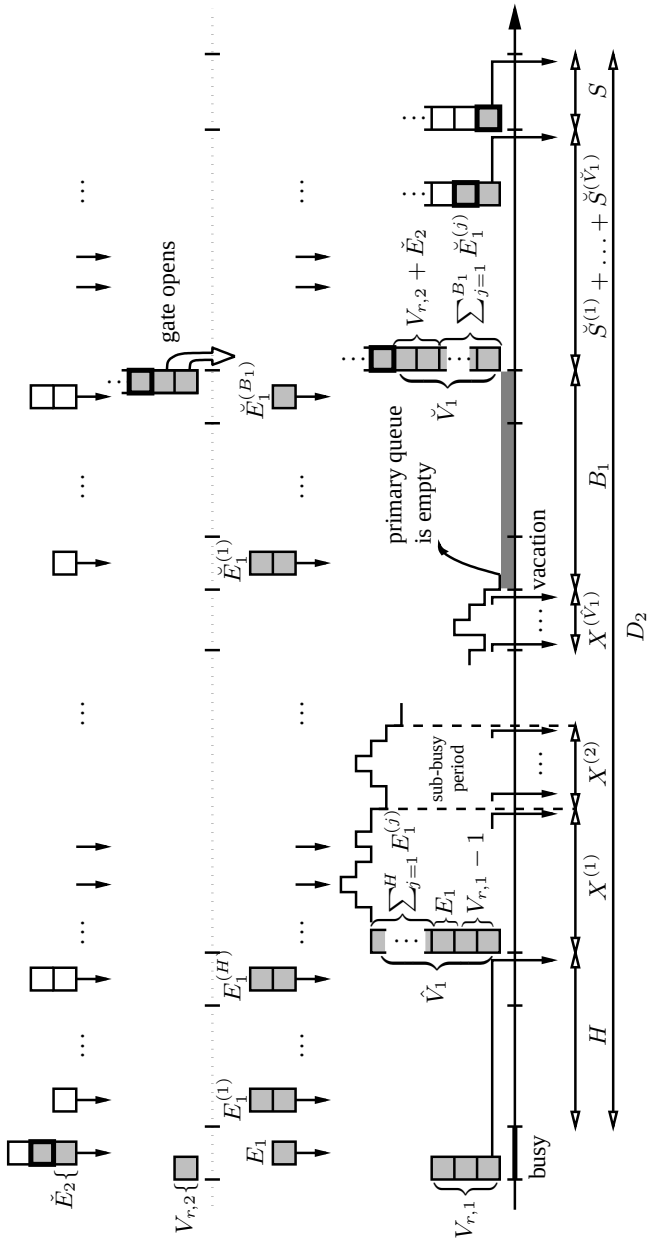


Figure 3.6: Delay experienced by a tagged customer arriving in the secondary queue during a busy slot.

queue during the remaining service time of the customer in service and minus the customer in service. That is,

$$\hat{V}_1 = V_{r,1} + E_1 - 1 + \sum_{j=1}^{H} E_1^{(j)}. \tag{3.42}$$

Recall that the sub-busy period of a primary customer denotes the number of slots it takes to reduce the primary queue size with a single customer from the moment this customer starts service. Therefore, a vacation starts after a number of slots equal to the sum of the sub-busy periods of all $\hat{V}_1$ customers present in the primary queue after the departure slot of the customer that is in service during the tagged customer's arrival slot.

The tagged customer's delay includes the time it takes before a vacation starts as well as the vacation itself. At the end of the vacation, the tagged customer moves to the primary queue. As the gate only opens at the end of the vacation, one sees that customers that arrive in the primary queue during the vacation are served before the tagged customer. Further, all customers that were present in the secondary queue upon arrival of the tagged customer are served before this customer as well. These observations then lead to the following expression for the customer delay D_2:

$$D_2 = H + \sum_{j=1}^{\hat{V}_1} X^{(j)} + B_1 + \sum_{j=1}^{\breve{V}_1} \breve{S}^{(j)} + S, \tag{3.43}$$

where

$$\breve{V}_1 = V_{r,2} + \breve{E}_2 + \sum_{j=1}^{B_1} \breve{E}_1^{(j)} \tag{3.44}$$

denotes the number of customers that are served after the vacation but before the tagged customer. Further, $X^{(j)}$ denotes the sub-busy period of the j-th customer in the primary queue after departure of the customer in service during the tagged customer's arrival slot, $\breve{E}_1^{(j)}$ denotes the number of primary arrivals in the j-th vacation slot following the tagged customer's arrival slot and $\breve{S}_j$ denotes the service time of the j-th customer served after the vacation following the tagged customer's arrival slot.

The former expressions and some standard z-transform manipulations then yield:

$$\begin{aligned} D_2(z) = {} & P_2(zE_1(S(z)), S(z), S(z))\, \breve{E}(S(z), S(z))\, S(z) \\ & + P_1(zE_1(X(z)), X(z), S(z))\, \breve{E}(X(z), S(z))\, B_1(zE_1(S(z)))\, S(z), \end{aligned} \tag{3.45}$$

in which $\breve{E}(z_1, z_2)$ denotes the joint probability generating function of the number of arrivals in the primary queue E_1 and those in the secondary queue that are served

before the tagged customer $\check{E}_2$ whereas $X(z)$ denotes the probability generating function of the sub-busy periods. We also used the fact that the state of the system at the beginning of a random customer's arrival slot and the state at a random slot boundary share the same stochastic properties. That is, equations (3.26) and (3.27) also denote the probability generating functions of the state of the system at the beginning of a random customer's arrival slot.

The generating function $\check{E}(z_1, z_2)$ follows from a similar argument as was used in section 1.2 to retrieve the probability generating function of the number of customers that arrive in the same slot as a random customer but that are served before this customer $\check{E}(z)$ (see e.g., Walraevens et al. [2002]),

$$\check{E}(z_1, z_2) = \frac{E(z_1, z_2) - E_1(z_1)}{\mu_{E_2}(z_2 - 1)}. \tag{3.46}$$

The probability generating function of the sub-busy period $X(z)$ of a primary customer in (3.45) is implicitly defined by

$$X(z) = S(z\, E_1(X(z))), \tag{3.47}$$

as the sub-busy period of a primary customer equals the sum of his service time and of the sub-busy periods of all arrivals in the primary queue during his service time (see e.g., Bruneel [1993]).

Plugging equation (3.26), (3.27) and (3.46) into (3.45), we finally retrieve,

$$\begin{aligned} D_2(z) = {} & \frac{(1-\rho)S(z)}{(\mu_{B_1} + \kappa\mu_{B_2})\mu_{E_2}(S(z)-1)} \cdot \\ & \left[\frac{E(S(z)) - E_1(S(z))}{zE_1(S(z)) - E(S(z))} \left\{ \begin{array}{l} \Omega(S(z))(B_1(zE_1(S(z))) - B_1(E(S(z)))) \\ + \kappa(B_2(zE_1(S(z))) - B_2(E(S(z)))) \end{array} \right\} \right. \\ & + B_1(zE_1(S(z))) \left\{ \begin{array}{l} B_1(E(X(z)))\Omega(X(z)) - \Omega(S(z)) \\ - \kappa(1 - B_2(E(X(z)))) \end{array} \right\} \\ & \left. \times \frac{S(zE_1(X(z))) - S(E(X(z), S(z)))}{X(z) - S(E(X(z), S(z)))} \frac{(E(X(z), S(z)) - E_1(X(z)))}{(zE_1(X(z)) - E(X(z), S(z)))} \right] \end{aligned} \tag{3.48}$$

The moment generating property of probability generating functions then yields various moments of the delay of customers arriving in the secondary queue.

3.1.6 Special cases

As stated, the gated-exhaustive model encapsulates the exhaustive and the gated vacation queueing systems with single and multiple vacations. We here verify this assertion and focus on the probability generating functions of the customer delay in these particular cases.

If there are only arrivals in the secondary queue ($E(z_1, z_2) = E(1, z_2)$), the system operates as a gated vacation queue. Only customers in the primary queue are served and customers move to the latter at the end of vacations. Therefore, no customers are served that arrived since the last vacation period in accordance with the gated vacation scheduling. On the other hand, if all customers arrive in the primary queue ($E(z_1, z_2) = E(z_1, 1)$), the system operates as an exhaustive vacation system. Clearly, there are no customers in the secondary queue in this case. Vacations therefore only start when there are no customers at all in the system.

For $B_1(z) = B_2(z)$, one observes that the system operates as a multiple vacation system. That is, as long as there are no customers in the system upon returning from a vacation, the server takes another vacation and all consecutive vacations share a common probability generating function. On the other hand, one obtains the single vacation model for $B_2(z) = z$. In this case, the server takes vacations of length 1 if there are no customers present in the system upon returning from a vacation. That is, the server checks for arrivals after each slot. As such, the system operates as a single vacation model.

Exhaustive vacation systems

We first focus on the exhaustive system. There are no customer arrivals in the secondary queue. Therefore, the delay of a random customer corresponds to the delay of a random customer in the primary queue. From equation (3.39), we obtain the probability generating functions of the customer delay for the exhaustive multiple and single vacation systems:

$$D^{(\text{multiple})}(z) = \frac{1-\rho}{\mu_E \mu_B} \frac{(1 - E(S(z)))(1 - B(z))S(z)}{(S(z) - 1)(z - E(S(z)))} \tag{3.49}$$

and

$$D^{(\text{single})}(z) = \frac{1-\rho}{\mu_E} \frac{(1 - E(S(z)))S(z)}{(S(z) - 1)(z - E(S(z)))} \frac{\left\{ \begin{array}{c} (1 - E(0))(1 - B(z)) \\ + B(E(0))(1 - z) \end{array} \right\}}{\mu_B(1 - E(0)) + B(E(0))}. \tag{3.50}$$

Here $B(z)$ and μ_B denote the common probability generating function and the mean length of a server vacation. Notice that the queue is empty before the start of the

vacation for exhaustive systems. Since $\Omega(z)$ is the probability generating function of the number of customers in the system before the start of the vacation, we find, $\Omega(z) = 1$. Equation (3.23) then implies that also κ is known. Therefore, in the case of exhaustive vacation systems, one does not need to determine the value κ numerically. The former expressions were also retrieved by Takagi [1993].

Gated vacation systems

For the gated vacation system, there are only arrivals in the secondary queue. Therefore, the delay of a random customer equals the delay of a random customer arrival in the secondary queue. Equation (3.48) yields the probability generating functions of the customer delay for the gated multiple and single vacation systems:

$$D^{(\text{multiple})}(z) = \frac{1-\rho}{\mu_E \mu_B} \frac{(1-E(S(z)))(1-B(z))S(z)}{(S(z)-1)(z-E(S(z)))} \Theta(S(z)) \tag{3.51}$$

and

$$D^{(\text{single})}(z) = \frac{(1-\rho)(E(S(z))-1)S(z)}{\mu_E(\mu_B+\kappa)} \frac{(z-1)\kappa + (B(z)-1)\Gamma(S(z))}{(z-E(S(z)))(S(z)-1)}. \tag{3.52}$$

Again, $B(z)$ and μ_B denote the common probability generating function and the mean length of a server vacation. Further, $\Theta(z)$ and $\Gamma(z)$ are defined as follows:

$$\Theta(z) \triangleq \frac{\Omega(z)+\kappa}{1+\kappa} B(E(z)), \tag{3.53}$$

$$\Gamma(z) \triangleq \Omega(z)B(E(z)) + \kappa(1-E(z)). \tag{3.54}$$

One can show that $\Theta(z)$ and $\Gamma(z)$ are the probability generating functions of the queue content at the end of a random vacation for the multiple vacation system and at the beginning of a busy period for the single vacation system respectively. For either generating function, equation (3.22) yields an implicit expression:

$$\Theta(z) = \Theta(S(E(z)))\, B(E(z)), \tag{3.55}$$

$$\Gamma(S(E(z))) = \frac{\Gamma(z) + \kappa(1-E(z))}{B(E(z))}. \tag{3.56}$$

From equations (3.51) and (3.55), one observes that for the gated multiple vacation system one again escapes numerical determination of the value κ. This is not the case for the gated single vacation system. In this case, the series y_i converges to κ for

increasing i:

$$\begin{cases} x_0 = 0, & \\ x_{i+1} = S(E(x_i)) & \text{for } i \geq 0, \\ y_0 = \dfrac{B(E(0))}{1 - E(0)}, & \\ y_{i+1} = \dfrac{B(E(x_{i+1}))\, y_i}{1 + (1 - E(x_{i+1}))\, y_i} & \text{for } i \geq 0. \end{cases} \tag{3.57}$$

These results are in accordance with those of Takagi [1991] and Fiems et al. [2004].

3.1.7 Numerical Examples

We now illustrate our results with some numerical examples.

We first concentrate on a multiple vacation system. The numbers of arrivals in the primary and the secondary queue during a slot are independent random variables that follow a Poisson distribution with mean values $\mu_{E_1} = x\mu_E$ and $\mu_{E_2} = (1 - x)\mu_E$ respectively. Here μ_E denotes the mean number of arrivals in the system in a slot and x denotes the fraction of all customers that arrive in the primary queue. Further, the service times of the customers share a common shifted geometrical distribution with mean $\mu_S = 5$ slots and the lengths of the vacations are deterministically equal to 20 slots. The reader is referred to appendix A for the characteristics of these distributions.

Figures 3.7 and 3.8 depict the mean and the variance of primary and secondary queue content versus the fraction x respectively. The latter are depicted for various values of the load $\rho = \mu_E \mu_S$ as indicated.

Clearly $x = 0$ corresponds to the purely gated vacation system, whereas $x = 1$ corresponds to the purely exhaustive vacation system. As the total load is fixed, an increase of x means that there are less customer arrivals in the secondary queue and more customer arrivals in the primary queue. One therefore expects that mean and variance of the secondary queue content decrease which is confirmed by the plots. The shapes of the curves of mean and variance of the primary queue content on the other hand are not easily explained.

Figure 3.9 depicts the correlation factor between primary and secondary queue content at random slot boundaries versus the fraction of customer arrivals x that arrive in the primary queue. The different curves correspond to different system loads ρ as depicted. The curves indicate a reasonable amount of correlation between both queue content and both positive and negative correlation are possible. For the purely gated vacation system ($x = 0$), correlation between primary and secondary queue content is negative. This follows from the fact that on the average the primary queue builds down while the secondary queue builds up. Further, for a purely exhaustive system

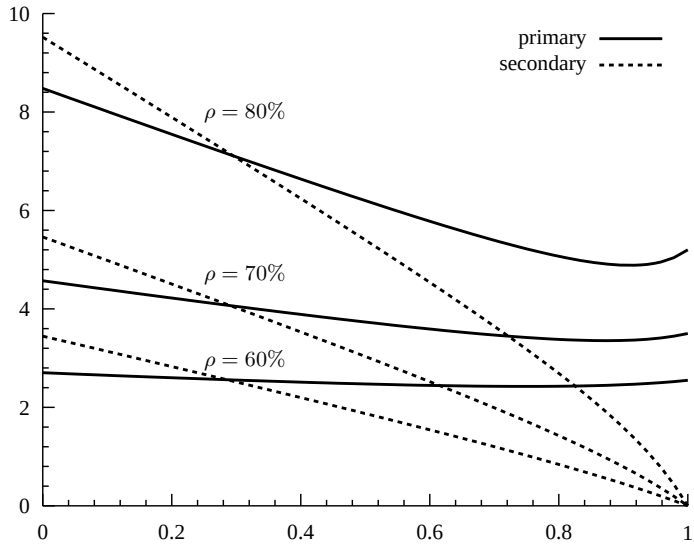


Figure 3.7: Mean primary and secondary queue content vs. the fraction x of customers arriving in the primary queue for different values of the total load ρ as indicated. (Poisson arrivals in both queues, shifted geometrically distributed service times with mean $\mu_S = 5$ slots, deterministic vacations of 20 slots.)

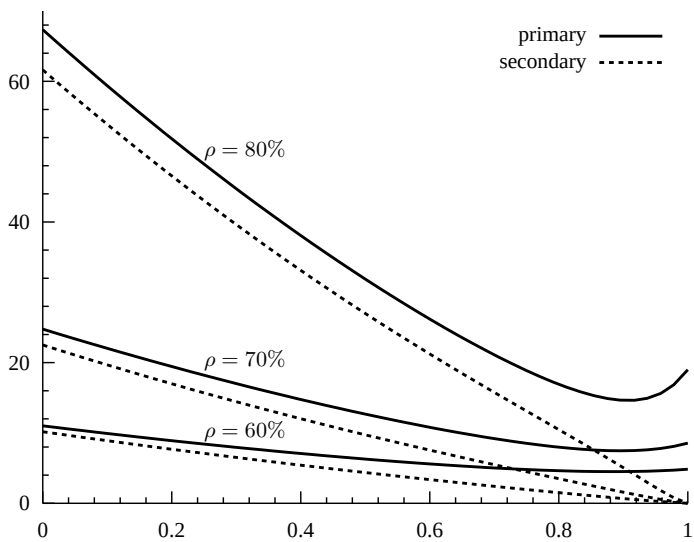


Figure 3.8: Variance of primary and secondary queue content vs. the fraction x of customers arriving in the primary queue for different values of the total load ρ as indicated. (Poisson arrivals in both queues, shifted geometrically distributed service times with mean $\mu_S = 5$ slots, deterministic vacations of 20 slots.)

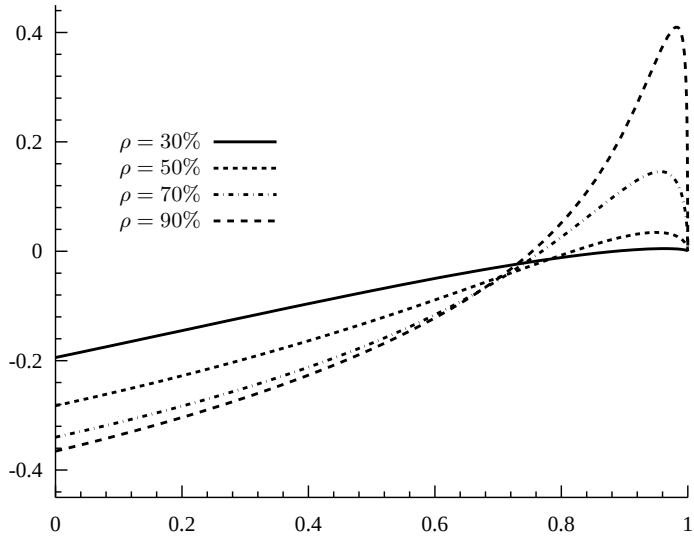


Figure 3.9: Correlation between the primary and secondary queue content vs. the fraction x of customers arriving in the primary queue for different values of the total load ρ as indicated. (Poisson arrivals in both queues, shifted geometrically distributed service times with mean $\mu_S = 5$ slots, deterministic vacations of 20 slots.)

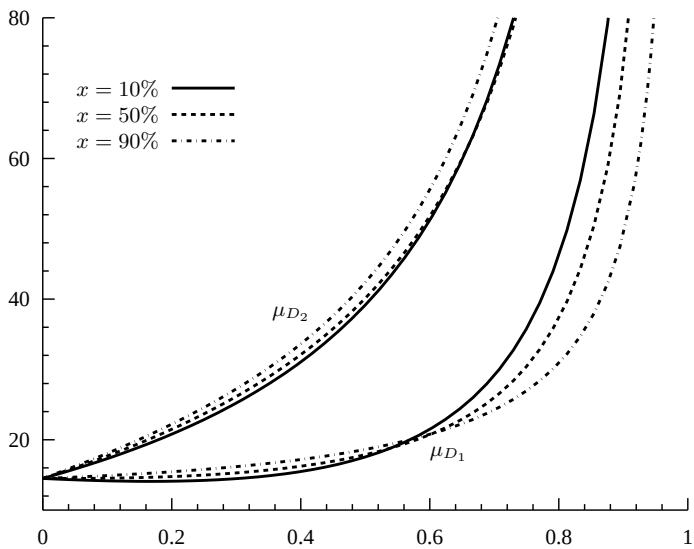


Figure 3.10: Mean delay of customers arriving in the primary (μ_{D_1}) and secondary (μ_{D_2}) queue vs. the load ρ for different values of the fraction x of customers that arrive in the primary queue. (Poisson arrivals in both queues, shifted geometrically distributed service times with mean $\mu_S = 5$ slots, deterministic vacations of 20 slots.)

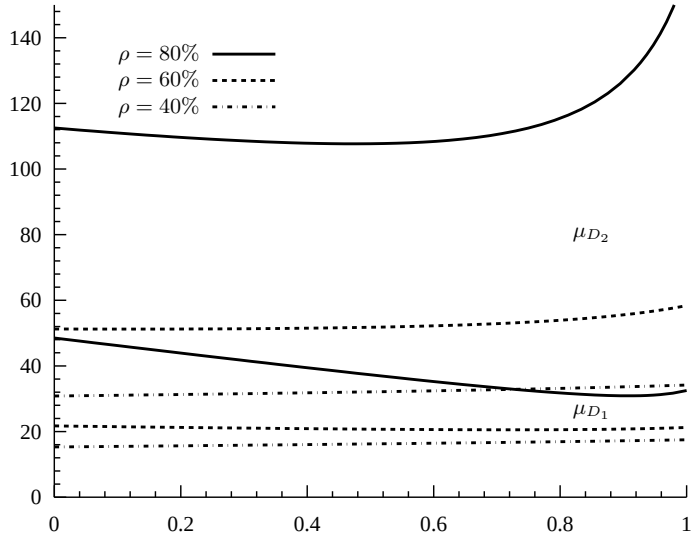


Figure 3.11: Mean delay of customers arriving in the primary (μ_{D_1}) and secondary (μ_{D_2}) queue vs. the fraction x of customers that arrive in the primary queue for different values of the total load ρ as indicated. (Poisson arrivals in both queues, shifted geometrically distributed service times with mean $\mu_S = 5$ slots, deterministic vacations of 20 slots.)

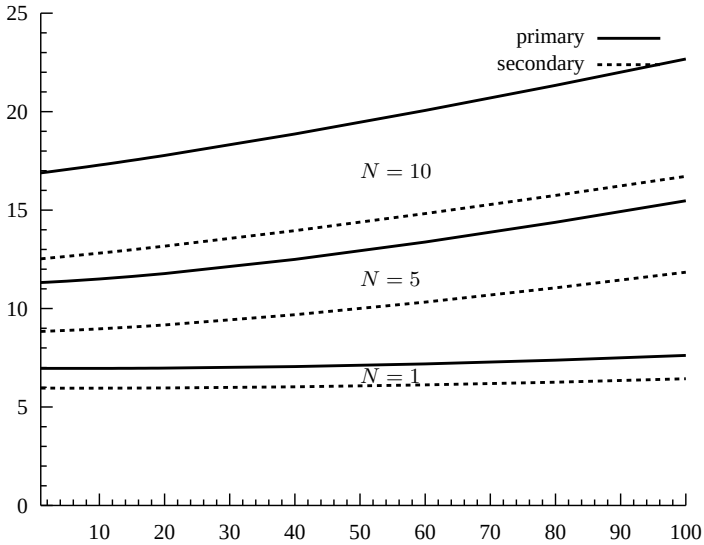


Figure 3.12: Mean primary and secondary queue content vs. the mean vacation length μ_{B_2} of a vacation immediately preceded by a vacation. (arrivals occur in deterministic batches of size N, shifted geometrical customer service times with mean $\mu_S = 5$ slots, shifted geometrical vacations, the mean length of the first vacation after a busy period equals 20 slots.)

($x = 1$), the correlation between primary and secondary queue content equals 0 as the secondary queue is always empty.

Figure 3.10 depicts mean customer delays μ_{D_1} and μ_{D_2} for customers arriving in primary and secondary queues (primary and secondary delay) respectively versus the total arrival load ρ for the multiple-vacation policy. Different values for the fraction of customers that arrive in the primary queue x are assumed as depicted. Increasing load implies longer delays and the mean primary delay is shorter than the mean secondary delay, as expected. Further, increasing the fraction x of customers that arrive in the primary queue may – depending on the total load – either increase or decrease the mean primary and/or the mean secondary delay.

This observation is also illustrated by figure 3.11 where we depict the mean customer delays μ_{D_1} and μ_{D_2} versus the fraction x of customers that arrive in the primary queue for different values of the total load as indicated. Depending on the load, one either obtains a performance gain or a performance loss by increasing the fraction of customers that arrive in the primary queue.

We conclude this section by considering a numerical example without multiple (identically distributed) vacations. The joint probability generating function of the numbers of arrivals in primary and secondary queue is given by,

$$E(z_1, z_2) = \left(1 - \frac{\mu_{E_1}}{N_1} + \frac{\mu_{E_1}}{N_1} z_1^{N_1}\right)\left(1 - \frac{\mu_{E_2}}{N_2} + \frac{\mu_{E_2}}{N_2} z_2^{N_2}\right), \tag{3.58}$$

with $0 \leq \mu_{E_i} \leq N_i$ ($i = 1, 2$). That is, during a slot, there are either N_1 (N_2) arrivals or no arrivals at all in the primary (secondary) queue. As before, the customer service times are shifted geometrically distributed with mean $\mu_S = 5$. Also the vacations follow a shifted geometrical distribution (with mean μ_{B_1} or μ_{B_2}).

Figure 3.12 depicts the mean primary and secondary queue content versus the mean vacation length of a vacation that is immediately preceded by a vacation. The arrival load is equally spread over primary and secondary queue ($\mu_{E_1} = \mu_{E_2}$) and different values of $N = N_1 = N_2$ are considered as depicted. Further, the mean length of a vacation that is not preceded by another vacation equals $\mu_{B_1} = 20$ slots. Performance is best for $N = 1$ and hardly influenced by the mean length of a vacation immediately preceded by another vacation. For higher N, performance deteriorates as higher N implies that more customers arrive at the same time. Also, the influence of the mean vacation length μ_{B_2} increases as the probability that the queue is empty upon returning from a vacation decreases for higher N. The latter follows from the fact that given a fixed arrival load, arrivals (of batches of customers) occur less frequent in time for higher N.

3.2 Non-gated vacation queues

In this section we consider a vacation process that encapsulates numerous classical non-gated vacation models such as exhaustive, time-limited and number-limited vacation systems. Our model's server leaves for a vacation at the end of a slot depending on the state of the system as well as on the (Markovian) state of the vacation process (see further). As in chapter 2, the server can leave for a vacation while a customer is in service. We again consider the CAI, RAI,wr and RAI operation modes to handle interrupted service. This section follows the lines of our contribution [Fiems and Bruneel, 2003].

3.2.1 Mathematical Model

Apart from the vacation process, we make approximately the same assumptions as in chapter 2. We consider a discrete-time queueing system with infinite storage capacity and a single server which may go on leave. The numbers of customers arriving during the consecutive slots constitute a series of i.i.d. non-negative random variables with common probability generating function $E(z)$ and the service times of these consecutive customers constitute a series of i.i.d. positive random variables with common probability generating function $S(z)$. We also assume that service times are bounded by some maximal value $S_{\max}$. Notice that this implies that $S(z)$ is a polynomial. We will relax the latter assumption where possible.

We now focus on the vacation process. Whenever the server is available during a slot (this slot is an A-slot), the vacation process is in 1 out of N possible states, say state 1 to N. At the end of an A-slot, the server leaves for a vacation during a number of slots and the vacation process returns to some other (or the same) state after this vacation. The vacation process is characterised by following probabilities:

- Given state i in a particular A-slot and given that there is a customer in service that does not end service during this slot, the server takes a vacation of n ($n \geq 0$) slots and goes to state j after this vacation with probability $b_1^{(ij)}(n)$.
- Similarly, given state i in a particular A-slot and given that a customer ends service in this slot and that the system is non-empty after departure of this customer, the server takes a vacation of n ($n \geq 0$) slots and goes to state j after this vacation with probability $b_2^{(ij)}(n)$.
- Also, given state i in a particular A-slot and given that a customer ends service in this slot and that the system is empty after departure of this customer, the server takes a vacation of n ($n \geq 0$) slots and goes to state j after this vacation with probability $b_3^{(ij)}(n)$.

- Finally, given state i in a particular A-slot and given that there are no customers in the system at the beginning of this slot, the server takes a vacation of n ($n \geq 0$) slots and goes to state j after this vacation with probability $b_4^{(ij)}(n)$.

Notice that we here allow zero-length vacations. The server then remains available.

Let $B_k^{(ij)}(z)$ ($k = 1 \ldots 4$) denote the partial conditional probability generating functions that correspond with the probabilities $b_k^{(ij)}(n)$, that is,

$$B_k^{(ij)}(z) = \sum_{n=0}^{\infty} b_k^{(ij)}(n) z^n. \tag{3.59}$$

The vacation process is then also characterised by the $N \times N$ matrices $\mathbf{B}_k(z)$ of these generating functions:

$$\mathbf{B}_k(z) = \left[B_k^{(ij)}(z) \right]_{i,j=1\ldots N}, \tag{3.60}$$

for $k = 1 \ldots 4$. To simplify notation, we further define $\tilde{\mathbf{B}}_k(z)$ as $z\,\mathbf{B}_k(z)$ for $k = 1 \ldots 4$.

3.2.2 Service completion times

We first focus on the probability generating functions of the service completion times for the different operation modes under consideration. Recall that a customer's service completion time starts at the beginning of the slot where the customer receives service for the first time and ends at the end of the slot where the customer leaves the system (see section 2.4).

Continue after interruption

We first consider the CAI operation mode. Let $c^{(ij)}(n|k)$ denote the probability that the service completion time of a customer takes n slots and that the server is in state j during the last slot of the service completion time, given that the server is in state i during the first slot of the service completion time, and given that this customer needs k slots service. Notice that – in accordance with the definition of the service completion time – the server is available during the first and last slot of a customer's service completion time.

Conditioning on the length of the vacation taken after the first service completion slot and the state of the server after this vacation, we get for $k > 1$ and for $i, j \in \{1 \dots N\}$,

$$c^{(ij)}(n|k) = \sum_{l=1}^{N} \sum_{m=0}^{n-k} c^{(lj)}(n-m-1|k-1)\, b_1^{(il)}(m), \tag{3.61}$$

for $n \geq k$ whereas the former probability equals 0 for $n < k$ as a customer's service completion time is at least as long as his service time. Equation (3.61) holds as from a system point of view, there is no difference between serving the remaining service time of a customer and serving a new customer with service time equal to that remaining service time.

Let $C^{(ij)}(z|k)$ denote the partial conditional probability generating function corresponding to the preceding probabilities, that is,

$$C^{(ij)}(z|k) \triangleq \sum_{n=1}^{\infty} c^{(ij)}(n|k)\, z^n \tag{3.62}$$

for $i, j \in \{1 \dots N\}$ and $k \geq 1$. Plugging equation (3.61) into the former expression, we retrieve following recursive expression for the partial conditional probability generating function $C^{(ij)}(z|k)$,

$$C^{(ij)}(z|k) = z \sum_{l=1}^{N} C^{(lj)}(z|k-1)\, B_1^{(il)}(z), \tag{3.63}$$

for $i, j \in \{1 \dots N\}$ and $k > 1$. Further, the service completion time of a customer that requires one slot of service equals one slot in accordance with the definition of the service completion times. As in this particular case, first and last slot of the customer's service completion time coincide, the state during the last slot equals the state during the first slot. Therefore we get,

$$C^{(ij)}(z|1) = \delta_{ij}\, z. \tag{3.64}$$

The Kronecker delta function δ_{ij} equals 1 for $i = j$ and equals 0 if this is not the case.

For ease of notation, let $\mathbf{C}(z|k)$ denote the $N \times N$ matrix with elements $C^{(ij)}(z|k)$ $(i, j = 1 \dots N)$, equation (3.63) then transforms into following matrix equation,

$$\mathbf{C}(z|k) = z\, \mathbf{B}_1(z) \cdot \mathbf{C}(z|k-1) = \tilde{\mathbf{B}}_1(z) \cdot \mathbf{C}(z|k-1), \tag{3.65}$$

for $k > 1$. In accordance with equation (3.64) we further find $\mathbf{C}(z|1) = z\, \mathbf{I}_N$. Here $\mathbf{I}_N$ denotes the $N \times N$ unity matrix. Successive application of equation (3.65) then

leads to

$$\mathbf{C}(z|k) = z\,\tilde{\mathbf{B}}_1(z)^{k-1}. \tag{3.66}$$

Let $c^{(ij)}(n)$ denote the probability that the service completion time of a customer takes n slots and that the server is in state j during the last slot of the service completion time, given that the server is in state i during the first slot of the service completion time. Further, $C^{(ij)}(z)$ denotes the partial conditional probability generating corresponding with this probability:

$$C^{(ij)}(z) \triangleq \sum_{n=1}^{\infty} c^{(ij)}(n)\, z^n, \tag{3.67}$$

for $i, j \in \{1 \ldots N\}$ and $\mathbf{C}(z)$ denotes the $N \times N$ matrix with elements $C^{(ij)}(z)$. Clearly, $\mathbf{C}(z)$ and $\mathbf{C}(z|k)$ relate as follows,

$$\mathbf{C}(z) = \sum_{k=1}^{S_{\max}} s(k)\mathbf{C}(z|k) = \sum_{k=1}^{S_{\max}} s(k) z\,\tilde{\mathbf{B}}_1(z)^{k-1}. \tag{3.68}$$

The matrix $\mathbf{C}(z)$ will be used in our further analysis.

Repeat after interruption with resampling

For RAI,wr we proceed similarly. Let $c^{(ij)}(n|k)$ ($n, k \geq 1$, $i, j \in \{1, \ldots, N\}$) denote the probability that the service completion time of a customer takes n slots and that the server is in state j during the last slot of this service completion time, given that the customer needs k slots of service (under the assumption that there are no vacations) and that the server is in state i during the first slot of the service completion time. Similarly, let $c^{(ij)}(n)$ denote the former probability without conditioning on the customer service time. Conditioning on the length of the vacation taken after the first effective service slot and the state of the server after this vacation we get for $n > 0$, for $k > 1$ and for $i, j \in \{1 \ldots N\}$,

$$c^{(ij)}(n|k) = \sum_{l=1}^{N} c^{(lj)}(n-1|k-1)\, b_1^{(il)}(0) + \sum_{l=1}^{N} \sum_{m=1}^{n-2} c^{(lj)}(n-m-1)\, b_1^{(il)}(m). \tag{3.69}$$

The former equality holds, as from a system point of view, there is no difference between serving a customer another time with a newly sampled service time and serving a new customer.

Let $C^{(ij)}(z|k)$ and $C^{(ij)}(z)$ denote the (partial conditional) probability generating functions corresponding to $c^{(ij)}(n|k)$ and $c^{(ij)}(n)$ respectively:

$$C^{(ij)}(z|k) = \sum_{n=1}^{\infty} c^{(ij)}(n|k)\, z^n, \tag{3.70}$$

$$C^{(ij)}(z) = \sum_{n=1}^{\infty} c^{(ij)}(n)\, z^n. \tag{3.71}$$

Equation (3.69) and some standard z-transform manipulations then yields for $k > 1$:

$$C^{(ij)}(z|k) = \sum_{l=1}^{N} z\, C^{(lj)}(z|k-1)\, B_1^{(il)}(0) + \sum_{l=1}^{N} z\, C^{(lj)}(z)\, (B_1^{(il)}(z) - B_1^{(il)}(0)). \tag{3.72}$$

Further, the service completion time of a customer that requires one slot of service equals one slot in accordance with the definition of the service completion times. As the first and the last slot of the customer's service completion time coincide in this case, we get,

$$C^{(ij)}(z|1) = \delta_{ij}\, z. \tag{3.73}$$

Similarly as for CAI, let $\mathbf{C}(z|k)$ and $\mathbf{C}(z)$ denote the $N \times N$ matrices with elements $C^{(ij)}(z|k)$ and $C^{(ij)}(z)$ respectively. Equations (3.72) and (3.73) can then be rewritten as,

$$\mathbf{C}(z|k) = z\, \mathbf{B}_1(0)\, \mathbf{C}(z|k-1) + z\, (\mathbf{B}_1(z) - \mathbf{B}_1(0))\, \mathbf{C}(z), \tag{3.74}$$

$$\mathbf{C}(z|1) = z\, \mathbf{I}_N, \tag{3.75}$$

for $k > 1$. Combining equation (3.75) and successive application of equation (3.74), we obtain

$$\begin{aligned}\mathbf{C}(z|k) &= z^k\, \mathbf{B}_1(0)^{k-1} \\ &\quad + (z\, \mathbf{B}_1(0) - \mathbf{I}_N)^{-1}\, ((z\, \mathbf{B}_1(0))^{k-1} - \mathbf{I}_N)\, z\, (\mathbf{B}_1(z) - \mathbf{B}_1(0))\, \mathbf{C}(z).\end{aligned} \tag{3.76}$$

The matrices $\mathbf{C}(z|k)$ and $\mathbf{C}(z)$ are related as,

$$\mathbf{C}(z) = \sum_{k=1}^{S_{\max}} s(k)\, \mathbf{C}(z|k). \tag{3.77}$$

Plugging equation (3.76) into the former equation and solving for the matrix $\mathbf{C}(z)$

then yields

$$\mathbf{C}(z) = \mathbf{\Theta}(z)\left(\mathbf{I}_N - (z\mathbf{B}_1(0) - \mathbf{I}_N)^{-1}(\mathbf{\Theta}(z) - z\mathbf{I}_N)(\mathbf{B}_1(z) - \mathbf{B}_1(0))\right)^{-1} \quad (3.78)$$

with

$$\mathbf{\Theta}(z) \triangleq \sum_{k=1}^{S_{\max}} s(k)\mathbf{B}_1(0)^{k-1}z^k. \quad (3.79)$$

As for CAI, the matrix $\mathbf{C}(z)$ will be used in our further analysis.

Repeat after interruption

For RAI, we may proceed in the same way as we did for CAI and RAI,wr. However, it is easier to base our analysis on the obtained results for RAI,wr. We follow a similar approach as in section 2.4.2.

We first consider fixed length service times. Clearly, RAI and RAI,wr then operate similarly. Or, equivalently, substitution of $S(z) = z^k$ – the probability generating function corresponding to fixed length service times of k slots – in (3.78), yields an expression for the matrix of the (partial conditional) probability generating functions of the service completion time of a customer for RAI operation, given that this customer's service time equals k slots:

$$\mathbf{C}(z|k) = \mathbf{\Theta}_k(z)\left(\mathbf{I}_N - (z\mathbf{B}_1(0) - \mathbf{I}_N)^{-1}(\mathbf{\Theta}_k(z) - z\mathbf{I}_N)(\mathbf{B}_1(z) - \mathbf{B}_1(0))\right)^{-1}, \quad (3.80)$$

with

$$\mathbf{\Theta}_k(z) = \mathbf{B}_1(0)^{k-1}z^k. \quad (3.81)$$

Summation over all possible service times with respect to their probabilities then yields the following expression for the matrix $\mathbf{C}(z)$ of the partial conditional probability generating functions of the service completion times in case of RAI operation,

$$\mathbf{C}(z) = \sum_{k=1}^{S_{\max}} s(k)\mathbf{\Theta}_k(z) \left(\mathbf{I}_N - (z\mathbf{B}_1(0) - \mathbf{I}_N)^{-1}(\mathbf{\Theta}_k(z) - z\mathbf{I}_N)(\mathbf{B}_1(z) - \mathbf{B}_1(0))\right)^{-1}. \quad (3.82)$$

The latter matrix will be used in our further analysis.

3.2.3 Queue content

We first consider the queue content at departure epochs. The state (in the Markovian sense) of the system at departure epochs is completely characterised by the number of customers in the queue and by the state of the vacation process. Let $V_d^{(k)}$ denote the queue content at the k-th departure epoch and let $Q^{(k)}$ denote the state of the vacation process during the slot where this customer leaves the system. Further, let $V_d^{(k)}(z,j)$ denote the partial probability generating function of the queue content at the kth departure epoch given that the vacation process is in state j during the departure slot of the departing customer:

$$V_d^{(k)}(z,j) \triangleq \mathrm{E}\left[z^{V_d^{(k)}}|Q^{(k)}=j\right] \Pr\left[Q^{(k)}=j\right], \tag{3.83}$$

for $j \in \{1 \dots N\}$. For ease of notation, we let $\mathbf{V}_d^{(k)}(z)$ denote the row vector with elements $V_d^{(k)}(z,j)$,

$$\mathbf{V}_d^{(k)}(z) \triangleq \left[V_d^{(k)}(z,1) \dots V_d^{(k)}(z,N)\right]. \tag{3.84}$$

We now relate queue content at the k-th and $(k+1)$-th departure epochs. Given that the queue is empty after departure of the k-th customer, a vacation characterised by the matrix $\mathbf{B}_3(z)$ is taken, followed by vacations characterised by the matrix $\mathbf{B}_4(z)$ until there is at least one customer in the queue upon returning from a vacation. A customer is then served which leaves the system after his service completion time. On the other hand, given that the k-th customer leaves a non-empty system behind, a vacation characterised by the matrix $\mathbf{B}_2(z)$ is taken, and the $(k+1)$-th customer is immediately served. The latter again leaves the system after his service completion time. Using some standard z-transform and matrix manipulations, one retrieves the following relation between the vectors $\mathbf{V}_d^{(k+1)}(z)$ and $\mathbf{V}_d^{(k)}(z)$,

$$\begin{aligned}\mathbf{V}_d^{(k+1)}(z) = &\left(\mathbf{V}_d^{(k)}(z) - \mathbf{V}_d^{(k)}(0)\right) \mathbf{B}_2(E(z)) \frac{1}{z} \mathbf{C}(E(z)) \\ &+ \mathbf{V}_d^{(k)}(0) \left(\mathbf{B}_3(E(z)) - \mathbf{B}_3(E(0))\right) \frac{1}{z} \mathbf{C}(E(z)) \\ &+ \mathbf{V}_d^{(k)}(0) \mathbf{B}_3(E(0)) \\ &\quad \sum_{i=0}^{\infty} \tilde{\mathbf{B}}_4(E(0))^i \left(\tilde{\mathbf{B}}_4(E(z)) - \tilde{\mathbf{B}}_4(E(0))\right) \frac{1}{z} \mathbf{C}(E(z)),\end{aligned} \tag{3.85}$$

which further simplifies into,

$$\mathbf{V}_d^{(k+1)}(z) = \left[\mathbf{V}_d^{(k)}(z) \mathbf{B}_2(E(z)) + \mathbf{V}_d^{(k)}(0) \mathbf{\Omega}(z)\right] \frac{1}{z} \mathbf{C}(E(z)) \tag{3.86}$$

with,

$$\begin{aligned}\mathbf{\Omega}(z) = &\ \mathbf{B}_3(E(0)) \left(\mathbf{I}_N - \tilde{\mathbf{B}}_4(E(0))\right)^{-1} \left(\tilde{\mathbf{B}}_4(E(z)) - \mathbf{I}_N\right) \\ &+ \mathbf{B}_3(E(z)) - \mathbf{B}_2(E(z)). \end{aligned} \tag{3.87}$$

Recall that the matrix $\mathbf{C}(z)$ is given by (3.68), (3.78) or (3.82) depending on the operation mode under consideration.

Under the assumption that the system under consideration reaches steady state, let

$$\mathbf{V}_d(z) = [V_d(z,1) \dots V_d(z,N)] \triangleq \lim_{k\to\infty} \mathbf{V}_d^{(k)}(z) \tag{3.88}$$

denote the vector of partial probability generating functions of the queue content at departure times in steady state. Equation (3.86) then easily yields

$$\mathbf{V}_d(z) = \mathbf{V}_d(0)\, \mathbf{\Omega}(z)\, \mathbf{C}(E(z)) \left(z\, \mathbf{I}_N - \mathbf{B}_2(E(z))\, \mathbf{C}(E(z))\right)^{-1}. \tag{3.89}$$

As partial probability generating functions are bounded for $|z| \leq 1$, every zero z_0 of the denominator of $V_d(z,k)$ ($k = 1 \dots N$) with $|z_0| \leq 1$ is also a zero of the numerator. This allows us – together with the normalisation condition $\sum_{j=1}^{N} V_d(1,j) = 1$ – to retrieve the unknown vector $\mathbf{V}_d(0)$.

In accordance with the definitions (3.83) and (3.84), the probability generating function of the queue content at departure epochs in steady state $V_d(z)$ is given by,

$$V_d(z) = \mathbf{V}_d(z)\, \mathbf{e}^T. \tag{3.90}$$

Here $\mathbf{e}^T$ denotes the $N \times 1$ column vector with all elements equal to 1.

Consider now the probability generating function $V_r(z)$ of the queue content at random slot boundaries. Due to the characteristics of the arrival process, we can relate the latter to the generating function $V_d(z)$ of the queue content at departure epochs as (see section 1.2),

$$V_r(z) = \mu_E \frac{1-z}{1-E(z)} V_d(z). \tag{3.91}$$

The moment generating property of probability generating functions then allows one to determine performance measures such as mean and variance of queue content in steady state.

3.2.4 Special cases I: systems without service interruptions

Our model encapsulates numerous "classical" vacation models with and without service interruptions. In this section, we focus on vacation systems without service interruptions, that is, on non-preemptive systems. Notice that for non-preemptive systems, the results for CAI, RAI or RAI,wr are the same as service is never interrupted.

Exhaustive vacation systems

In a system with exhaustive vacations (see section 3.1.6), the server starts a vacation whenever the queue is empty after departure of a customer. If the queue is still empty upon returning from a vacation, the server either immediately takes another vacation or remains idle until a new customer arrives. One refers to these two policies as the multiple and the single vacation policy respectively. We can assess performance of the $Geo^X/G/1$ queue with multiple and single vacations using the following 1×1 vacation matrices:

$$\mathbf{B}_1(z) = \mathbf{B}_2(z) = \Big[1\Big], \tag{3.92}$$

$$\mathbf{B}_3(z) = \Big[B(z)\Big], \tag{3.93}$$

and using either $\mathbf{B}_4(z) = [1]$ or $\mathbf{B}_4(z) = [B(z)/z]$ for the single and the multiple vacation system respectively. Here, we assume that the consecutive vacation lengths constitute a series of i.i.d. positive random variables with common probability generating function $B(z)$.

As no vacations start during a customer's service time, the latter is never interrupted and therefore a customer's service completion time equals his service time ($\mathbf{C}(z) = [S(z)]$), a result that can also be found from (3.68), (3.78) or (3.82). One now easily observes that there is no need to assume that service times are bounded in this particular case.

Plugging the former matrices into our results – equations (3.89) to (3.91) – we retrieve the probability generating function of the queue content at random slot boundaries for the exhaustive multiple and single vacation systems:

$$V_r^{(\text{multiple})}(z) = (1-\mu_E\mu_S)\frac{(z-1)S(E(z))}{z-S(E(z))}\,\frac{B(E(z))-1}{\mu_B\,(E(z)-1)} \tag{3.94}$$

$$\begin{aligned} V_r^{(\text{single})}(z) =& (1-\mu_E\mu_S)\frac{(z-1)S(E(z))}{z-S(E(z))} \\ &\times \frac{(1-E(0))\,(B(E(z))-1)+B(E(0))\,(E(z)-1)}{(E(z)-1)\,(\mu_B(1-E(0))+B(E(0)))} \end{aligned} \tag{3.95}$$

Notice that determination of the unknown vector $\mathbf{V}_d(0)$ here reduces to the determination of a single constant which follows from the normalisation condition. Our results comply with Takagi's results [Takagi, 1993, pg. 98 and pg. 132].

Exhaustive number-limited vacation systems

For exhaustive number-limited vacation systems (E-limited systems) the server takes a vacation whenever there are no more customers to be served or whenever a fixed number N of customers have been served since the last vacation. In case of the E-limited multiple vacation system, the server immediately leaves for another vacation if it finds an empty queue upon returning from a vacation. We can assess performance of the $Geo^X/G/1$ E-limited multiple vacation queueing system with the following $N \times N$ vacation matrices:

$$\mathbf{B}_1(z) = \mathbf{I}_N, \tag{3.96}$$

$$\mathbf{B}_2(z) = \mathbf{B}_3(z) = \left[\delta_{i(j-1)} + \delta_{iN}\delta_{j1}B(z)\right]_{i,j=1\ldots N}, \tag{3.97}$$

$$\mathbf{B}_4(z) = B(z)/z[\delta_{j1}]_{i,j=1\ldots N}. \tag{3.98}$$

Again, we assume that the consecutive vacation lengths constitute a series of i.i.d. positive random variables with common probability generating function $B(z)$. The state of the vacation process here corresponds to the numbers of customers that started service since the last vacation.

As no vacations start during a customer's service, the latter is never interrupted and therefore a customer's service completion time equals his service time. As the vacation state does not change during service we then find: $\mathbf{C}(z) = S(z)\mathbf{I}_N$. As for exhaustive vacation systems, one easily observes that there is no need to assume that service times are bounded.

Plugging the former matrices into our results – equations (3.89) to (3.91) – we obtain the probability generating function of the queue content at random slot boundaries:

$$V_r(z) = \frac{(N(1-\mu_E\mu_S)-\mu_E\mu_B)(1-z)(1-B(E(z)))S(E(z))z^{N-1}\Theta(z)}{\mu_B(1-E(z))(S(E(z))^N B(E(z)) - z^N)}. \tag{3.99}$$

The unknown function $\Theta(z)$ is a normalised ($\Theta(1) = 1$) polynomial of order $N-1$. Recall that probability generating functions are bounded within the unit disk. We can therefore determine the $N-1$ unknown coefficients of $\Theta(z)$ through the $N-1$ zeros of the denominator (more precisely of the factor $S(E(z))^N B(E(z)) - z^N$) of the former expression within the unit disk. The results comply with those of Takagi [Takagi, 1993, pp. 209 – 214].

Non-preemptive time-limited vacation systems

In time-limited systems, the server takes a vacation whenever there are either no more customers to be served or whenever a timer (restarted after each vacation) expires. In the particular case of non-preemptive time-limited multiple vacation system, the server takes a vacation after finishing a customer's service during which the timer expired. Further, the server immediately leaves for another vacation if it finds an empty queue upon returning from a vacation.

If one assumes a non-preemptive time-limited $Geo^X/G/1$ system with geometrically distributed timers, one retrieves this system's performance measures using the following matrices:

$$\mathbf{B}_1(z) = \begin{bmatrix} \alpha & 1-\alpha \\ 0 & 1 \end{bmatrix}, \tag{3.100}$$

$$\mathbf{B}_2(z) = \mathbf{B}_3(z) = \begin{bmatrix} \alpha + (1-\alpha)B(z) & 0 \\ B(z) & 0 \end{bmatrix}, \tag{3.101}$$

$$\mathbf{B}_4(z) = \begin{bmatrix} \frac{B(z)}{z} & 0 \\ 1 & 0 \end{bmatrix}. \tag{3.102}$$

Here, α denotes the probability that the timer does not expire during a slot and $B(z)$ denotes the common probability generating function shared by the consecutive (independent) vacations. Note that state 2 corresponds to slots where the timer is expired, whereas state 1 corresponds to slots where this is not the case.

As no vacations start during a customer's service, the latter is never interrupted and therefore a customer's service completion time equals his service time. We easily obtain the service completion time matrix by plugging equation (3.100) into either (3.68), (3.78) or (3.82):

$$\mathbf{C}(z) = \begin{bmatrix} \frac{S(\alpha z)}{\alpha} & S(z) - \frac{S(\alpha z)}{\alpha} \\ 0 & 1 \end{bmatrix}. \tag{3.103}$$

Again, one easily observes that there is no need to assume that service times are bounded.

Substitution of the former matrices in our results, leads to the following expression

for probability generating function of the queue content at random slot boundaries:

$$V_r(z) = \frac{[1-\mu_B\mu_E(1-S(\alpha))-\mu_S\mu_E]S(E(z))(z-1)(1-B(E(z)))}{\mu_B(B(E(z))S(E(z))+S(\alpha E(z))(1-B(E(z)))-z)(E(z)-1)}. \tag{3.104}$$

The results comply with those presented in Fiems and Bruneel [2001].

3.2.5 Special cases II: systems with service interruptions

The system under consideration also encapsulates some "classical" vacation systems where vacations can preempt customers in service.

Random vacations

Under the assumption that all $\mathbf{B}_i(z)$ ($i = 1\ldots4$) are equal, the vacations occur independently of the state of the system. In this case, the system under consideration reduces to a system with random vacations (see chapter 2). In particular, the system under consideration reduces to the $Geo^X/G/1$ vacation system with geometrically distributed A-times and generally distributed B-times under consideration in section 2.4 for

$$\mathbf{B}_i(z) = \Big[\alpha + (1-\alpha)B(z)\Big], \tag{3.105}$$

for $i = 1\ldots4$. Here, α denotes the probability that an A-period continues after some A-slot and $B(z)$ the probability generating functions of the B-periods or vacations. Plugging these matrices in our results we find,

$$\begin{aligned} V_r(z) = {} & \frac{1-\mu_E(\mu_C+(1-\alpha)\mu_B)}{1+(1-\alpha)\mu_B} \\ & \times \frac{C(E(z))(z-1)}{E(z)-1}\,\frac{1-\alpha\,E(z)-(1-\alpha)\,E(z)\,B(E(z))}{C(E(z))(\alpha+(1-\alpha)B(E(z)))-z}. \end{aligned} \tag{3.106}$$

with

$$C(z) = \frac{S(\alpha z+(1-\alpha)zB(z))}{\alpha+(1-\alpha)B(z)}, \tag{3.107}$$

$$C(z) = \frac{S(\alpha z)(1-\alpha z)}{\alpha(1-\alpha z)-(1-\alpha)B(z)(\alpha z-S(\alpha z))}, \tag{3.108}$$

and

$$C(z) = \sum_{k=1}^{\infty} s(k) \frac{(\alpha z)^k (1-\alpha z)}{\alpha(1-\alpha z) - (1-\alpha)B(z)\alpha z(1-(\alpha z)^{k-1})} \tag{3.109}$$

for CAI, RAI,wr and RAI operation modes respectively. Again, there is no need to have an upper bound for the customer service times as we can retrieve explicit expressions for the moments of the server completion times (see section 2.4). One now easily verifies that these results correspond to the results found in section 2.4.

Preemptive time-limited vacation systems

The server of a preemptive time-limited multiple vacation system leaves for a vacation when there are no more customers in the system or when a timer expires. The latter is restarted after each vacation. After the vacation, the interrupted customers either resume or repeat (with or without resampling) their service. Further, the server leaves immediately for another vacation upon returning from a vacation if there are no customers present in the system at that epoch.

The system under consideration reduces to a preemptive time-limited $Geo^X/G/1$ multiple vacation system with geometrically distributed timers using the following set of matrices:

$$\mathbf{B}_1(z) = \mathbf{B}_2(z) = \mathbf{B}_3(z) = \Big[\alpha + (1-\alpha)B(z)\Big], \tag{3.110}$$

$$\mathbf{B}_4(z) = \left[\tfrac{B(z)}{z}\right]. \tag{3.111}$$

Plugging the former expressions into our results, then leads to the following expression for the probability generating function of the queue content at random slot boundaries:

$$V_r(z) = \frac{[1-\mu_E(\mu_C + (1-\alpha)\mu_B)](1-B(E(z)))(z-1)C(E(z))}{\mu_B(E(z)-1)[C(E(z))(\alpha + (1-\alpha)B(E(z))) - z]} \tag{3.112}$$

Here, $C(z)$ denotes the probability generating function of a customer's service completion time. The latter – note that the matrix $\mathbf{B}_1(z)$ is the same as for the system with random vacations studied before – are given by (3.107), (3.108) and (3.109) for the preemptive resume (CAI), the preemptive repeat with resampling (RAI,wr) and the preemptive repeat (RAI) time-limited systems respectively.

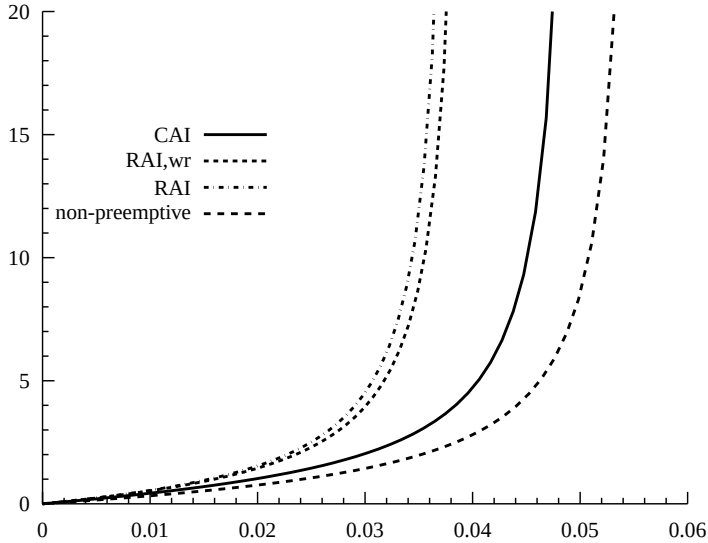


Figure 3.13: The mean queue content vs. the mean number of arrivals in a slot for the time-limited vacation system. (Poisson arrivals, shifted geometrically distributed vacations with mean $\mu_B = 20$ slots, a mean timer length of 20 slots, shifted symmetrical binomial customer service times with mean $\mu_S = 10$ slots.)

3.2.6 Numerical example

To illustrate our analysis, we now compare the different time-limited vacation systems numerically. That is, we compare the non-preemptive time-limited and the preemptive time-limited (CAI, RAI,wr and RAI) multiple-vacation systems with geometrically distributed timers.

We assume a Poisson arrival process. That is, the numbers of customers arriving in a random slot follow a Poisson distribution. Further, service times are shifted symmetrically binomially distributed whereas vacations follow a shifted geometrical distribution. The reader is referred to appendix A for the characteristics of these distributions.

Figure 3.13 depicts the mean queue content at random slot boundaries versus the mean number of arrivals in a slot. The mean timer length and the mean vacation length equal $\mu_A = \mu_B = 20$ slots whereas the mean customer service time equals $\mu_S = 10$ slots. For increasing load, the mean queue content increases as expected. Further, the non-preemptive time-limited vacation system outperforms the preemptive time-limited vacations systems. Due to service repetitions, the preemptive resume (CAI) time-limited system performs better than the preemptive repeat with resampling (RAI,wr) time-limited system which in turn outperforms the preemptive resume (RAI) time-limited system.

Figure 3.14 depicts the mean queue content at random slot boundaries versus the mean

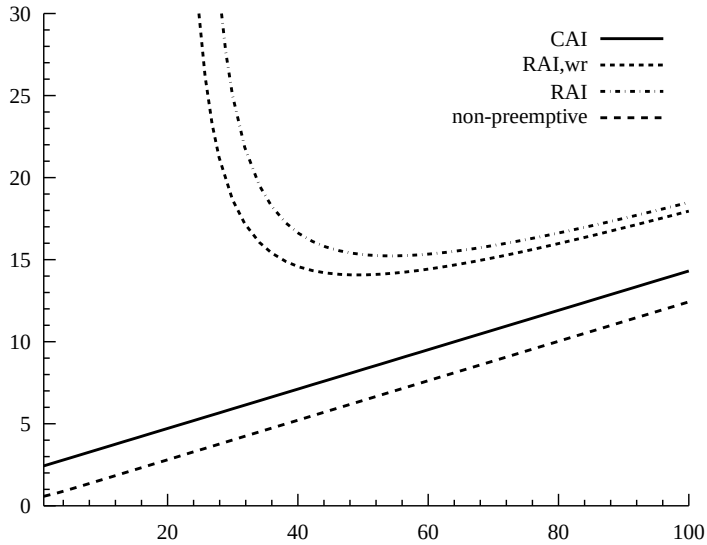


Figure 3.14: The mean queue content vs. the mean timer length for the time-limited vacation system. (Poisson arrivals, mean number of arrivals per slot of $\mu_E = 0.04$, shifted symmetrical binomial customer service times with mean $\mu_S = 10$ slots, shifted geometrical vacations, the mean vacation length equals the mean timer length.)

timer length. Again we assume that the mean customer service time equals $\mu_S = 10$ slots. The mean number of arrivals in a slot equals 0.04 and the mean vacation length equals the mean timer length. For both the non-preemptive timer and the preemptive resume timer, the mean queue content increases for increasing mean timer lengths. This comes from the fact that also the mean vacation lengths increase. That is, the queue builds up during longer vacation periods and builds down again during longer timer periods resulting in longer mean queue lengths. This is also the case for the preemptive repeat with and without resampling timers if the mean timer period is sufficiently large. As for the model with random vacations (see section 2.4) however, we observe that the mean timer length should be sufficiently large in case of preemptive time-limits with repetitions (RAI,wr and RAI). If the mean timer length is too small, there will be more and more service repetitions resulting in longer service completion times and therefore also in longer mean queue lengths.

Chapter 4

Summary

To conclude this dissertation, we now summarise our main contributions.

4.1 Random service vacations

In chapter 2, we investigated discrete-time queueing systems with random vacations. We first surveyed literature on both continuous-time and discrete-time queueing systems with random vacations. We then investigated the $Geo^X/G/1$ queueing system with three different vacation processes: the Bernoulli vacation process (sections 2.1, 2.2 and 2.6), the Markovian vacation process (section 2.3) and the on/off process with generally distributed vacation times (section 2.4). All analyses follow the probability generating functions approach and focus on queue content at different epochs in steady state and on customer delay.

Due to the combination of multi-slot customer service times and random service vacations – the server can leave for a vacation while there is a customer in service – we focused on different operation modes to cope with interrupted customer service times. In the continue after interruption mode, the interrupted customer resumes service after the interruption. On the other hand, in the repeat after interruption mode and the repeat after interruption mode with resampling, service has to start all over from the beginning. In the former mode, the service time after the interruption is the same as the service time before the interruption. In the latter mode, the service time is resampled. These three operation modes are investigated for all vacation processes under consideration.

The introduction of the concept of the effective service times simplified the queueing analysis considerably. This concept allows one to carry out a major part of the queueing

analysis simultaneously for all operation modes under consideration. To compare the method of the effective service times with the well known supplementary variable approach, we also presented a supplementary variable analysis for the model with Bernoulli vacations (section 2.2).

The theoretical results are applied to analyse a discrete-time multi-class preemptive priority system (section 2.5). We showed that the queueing model with the on/off process with generally distributed vacation times can be used to assess performance of a $Geo^X/G/1$ multi-class preemptive priority system exactly. The continue after interruption mode corresponds to the preemptive resume scheduling and the repeat after interruption mode (with resampling) corresponds to the preemptive repeat scheduling discipline. Numerically, we found that performance assessment of this type of priority systems by means of the Markovian model leads to fairly accurate results. On the other hand, performance analysis of priority systems by means of the Bernoulli model did not yield satisfying results.

We did not limit our discussion to the former three operation modes. For the Bernoulli vacation process, we also considered some variants (section 2.6). For the delayed modes (delayed RAI and delayed RAI,wr), service continues during vacations and is repeated (with or without resampling) until the customer receives service without vacations during his service time. For partial modes (p-RAI, p-RAI,wr, dp-RAI, dp-RAI,wr), a customer's service time is split up in parts and the former operation modes are not applied on the complete service time but on these service parts instead.

4.2 Other vacation models

In chapter 3, we investigated some other (non-random) vacation models. We first surveyed recent work on exhaustive, gated and time-limited vacation systems. We then proposed and investigated the gated-exhaustive vacation system. This model encapsulates both the gated and the exhaustive vacation systems with single and multiple vacations. The supplementary variable approach was applied to obtain various performance measures such as the probability generating functions of the queue content at various epochs in steady state and of the customer delay. Our results however depend on a variable κ which one has to determine numerically. We derived a recursion to obtain the latter variable.

Finally, we proposed and analysed a fairly general vacation model that encapsulates various queueing systems with non-gated vacations. The vacation process can capture correlation between consecutive available and vacation periods and can also – to some degree – take the state of the queueing system into account. As the model again allows that the server leaves for a vacation while a customer is in service, the CAI, RAI and RAI,wr operation modes were considered. The analysis methodology is similar to the effective service time approach followed in chapter 2 and we obtained the probability generating functions of the queue content at departure epochs and at random slot

boundaries. Matrices containing partial conditional probability generating functions were employed to cope with the finite state space of the vacation process. We further showed that the model encapsulates amongst others the exhaustive vacation system with single and multiple vacations, preemptive and non-preemptive time-limited vacation systems, systems with random vacations and number-limited vacation systems.

Appendix A

Frequently used distributions

We describe some properties of the various discrete distributions, encountered in the numerical examples throughout this dissertation. We focus in particular on the probability mass function, the probability generating function, the mean and the variance of these distributions.

Deterministic distribution The discrete deterministic distribution takes an integer value K with probability 1. Its probability mass function $\theta(n)$ and probability generating function $\Theta(z)$ are therefore given by,

$$\theta(n) = \begin{cases} 1 & \text{for } n = K, \\ 0 & \text{elsewhere,} \end{cases} \tag{A.1}$$

and,

$$\Theta(z) = z^K, \tag{A.2}$$

respectively. The mean value μ equals K whereas the variance σ^2 equals 0. Clearly, the distribution is completely characterised by its mean which must be integer.

Bernoulli distribution The Bernoulli distribution takes the value 1 with probability p and the value 0 with probability $1 - p$. Its probability mass function $\theta(n)$ and probability generating function $\Theta(z)$ are therefore given by,

$$\theta(n) = \begin{cases} 1 - p & \text{for } n = 0, \\ p & \text{for } n = 1, \\ 0 & \text{elsewhere,} \end{cases} \tag{A.3}$$

and,

$$\Theta(z) = 1 - p + p\,z, \tag{A.4}$$

respectively. Mean μ and variance σ^2 are given by p and $p(1-p)$ respectively. The distribution is completely characterised by its mean which must take a value between 0 and 1.

(Shifted) Geometrical distribution The geometrical distribution is the distribution of the number of trials until success in consecutive Bernoulli experiments. The shifted geometrical distribution is the distribution of the sum of a fixed integer (the shift) and a geometrically distributed random variable. Its probability mass function $\theta(n)$ and probability generating function $\Theta(z)$ are given by,

$$\theta(n) = \begin{cases} (1-\beta)\,\beta^{n-K} & \text{for } n \geq K, \\ 0 & \text{elsewhere,} \end{cases} \tag{A.5}$$

and,

$$\Theta(z) = z^K\,\frac{1-\beta}{1-\beta\,z}, \tag{A.6}$$

respectively. Here K denotes the shift and β denotes the probability that there is no success in a single Bernoulli trial. Mean μ and variance σ^2 are given by,

$$\mu = K + \frac{\beta}{1-\beta},$$
$$\sigma^2 = \frac{\beta}{(1-\beta)^2}.$$

The shifted geometrical distribution with shift K is characterised by this shift and its mean value μ which has to be larger than or equal to the shift. If we refer to the shifted geometrical distribution without specifying the shift, we assume that the latter equals 1. A geometrical distribution is a shifted geometrical distribution with shift 0. Note that in the latter cases, the distribution is completely characterised by its mean.

(Shifted) binomial distribution The binomial distribution corresponds to the distribution of the number of successes in a fixed number of Bernoulli trials. As for the geometrical distribution, we also consider the shifted variant. The shifted binomial distribution is the distribution of the sum of a fixed integer (the shift) and a binomially distributed random variable. Its probability mass function $\theta(n)$ and its probability

generating function $\Theta(z)$ are given by,

$$\theta(n) = \begin{cases} \binom{N}{n-K} (1-p)^{N+K-n}\, p^{n-K} & \text{for } K \leq n \leq K+N, \\ 0 & \text{elsewhere,} \end{cases} \tag{A.7}$$

and,

$$\Theta(z) = z^K\, (1 - p + p\, z)^N, \tag{A.8}$$

respectively. Here p denotes the probability of success of a single Bernoulli trial, N denotes the number of Bernoulli trials and K denotes the shift. The distribution is symmetrical for $p = \frac{1}{2}$. Mean and variance are given by,

$$\mu = K + N\, p,$$
$$\sigma^2 = N\, p\, (1-p).$$

The shifted binomial distribution is either characterised by the triple (K, N, p) where p takes a value between 0 and 1 and where K and N are non-negative integers or by the triple (K, N, μ) where μ takes a value between K and $K + N$ and where K and N are non-negative integers. If we refer to the shifted binomial distribution without specifying the shift K, we assume that the latter equals 1. A binomial distribution is a shifted binomial distribution with shift $K = 0$.

(Shifted) Poisson distribution The Poisson distribution is the distribution of the number of arrivals of a Poisson process during a time unit. As for the former distributions, we also consider the shifted variant. Its probability mass function $\theta(n)$ and its probability generating function $\Theta(z)$ are given by,

$$\theta(n) = \begin{cases} \dfrac{\lambda^{n-K}}{(n-K)!} \exp(-\lambda) & \text{for } n \geq K, \\ 0 & \text{elsewhere,} \end{cases} \tag{A.9}$$

and,

$$\Theta(z) = z^K\, \exp(\lambda\, (z-1)), \tag{A.10}$$

respectively. Here K denotes the shift and λ denotes the traffic intensity of the Poisson process. Mean μ and variance σ^2 of the shifted Poisson distribution are given by,

$$\mu = K + \lambda,$$
$$\sigma^2 = \lambda.$$

The shifted Poisson distribution is characterised by the (non-negative) integer shift K and by either the intensity λ or the mean μ. The latter must exceed K. Again we assume that the shift equals $K = 1$ if we refer to the shifted Poisson distribution without specifying the shift. The Poisson distribution is a shifted Poisson distribution with zero shift, that is, with $K = 0$.

Appendix B

From discrete to continuous time

As mentioned in chapter 1, the obtained probability generating functions of queue content and customer delay for the discrete-time queueing system can be used to retrieve corresponding results for the equivalent continuous-time queueing model. We will illustrate this assertion here. In particular, we relate the $Geo^X/G/1$ queueing system under investigation in section 1.2 to its continuous-time counterpart: the $M^X/G/1$ system.

B.1 The continuous-time model

The $M^X/G/1$ queueing model is characterised as follows. During time, arrivals occur in batches in an infinite capacity buffer. There is a single server which serves customers in order of arrival. The time between consecutive batches follows an exponential distribution with mean $1/\lambda$ and the size (the number of customers) of the consecutive batches constitutes a series of i.i.d. positive random variables with common probability generating function $\Psi(z)$. That is, we consider a batch Poisson arrival process with batch arrival intensity λ and batch size probability generating function $\Psi(z)$.

Further, service times of the consecutive customers constitute a series of i.i.d. random variables with common density function $s(t)$ and corresponding Laplace transform $\mathcal{S}(x)$,

$$\mathcal{S}(x) = \int_0^\infty s(t)\exp(-xt)dt. \tag{B.1}$$

B.2 Adapting arrival and service processes

In a first step, we adapt the arrival process and the service process of the discrete-time $Geo^X/G/1$ system such that for decreasing slot lengths Δ, these converge to a batch Poisson process and a continuous-time independent service process respectively. That is, for decreasing slot lengths, we obtain an $M^X/G/1$ queueing system as described in the preceding section. The adaptation yields slot length dependent expressions for the probability generating functions of the arrival and service processes.

Remember that the exact timing within slots does not influence discrete-time performance measures such as queue content at departure epochs and at random slot boundaries and customer delay. We can therefore assume that all arrivals in and departures from the discrete-time $Geo^X/G/1$ queueing system occur at the end of slots, i.e., just before slot boundaries.

Let $E(z|\Delta)$ denote the probability generating function of the number of arrivals in a slot for the discrete-time system, given slotlength Δ. We make the discrete-time arrival process Δ-dependent as follows,

$$E(z|\Delta) \triangleq 1 - \lambda\,\Delta + \lambda\Delta\,\Psi(z). \tag{B.2}$$

Here λ and $\Psi(z)$ are the characteristics of the continuous-time arrival process.

It is now easy to show that for decreasing Δ, the arrival process converges to a Poisson batch arrival process characterised by λ and $\Psi(z)$. Remember that the time (in slots) between consecutive batches of arrivals in the discrete-time system follows a geometrical distribution. The common probability generating function $I(z|\Delta)$ is given by,

$$I(z|\Delta) = \frac{(1 - E(0|\Delta))\,z}{1 - E(0|\Delta)\,z} = \frac{\lambda\Delta z}{1 - (1 - \lambda\Delta)z}, \tag{B.3}$$

whereas the probability generating function of the batch sizes is given by,

$$\Psi(z|\Delta) = \frac{E(z|\Delta) - E(0|\Delta)}{1 - E(0|\Delta)} = \Psi(z). \tag{B.4}$$

The former equalities simplify as $\Psi(0) = 0$, i.e., there is at least one customer in a batch.

Consider now the Laplace transform $\mathcal{I}(x|\Delta)$ of the interarrival times in absolute time. Due to the timing assumptions within slots, the interarrival time in absolute time equals the product of the interarrival time in slots and the slotlength. Therefore, the latter Laplace transform is given by,

$$\mathcal{I}(x|\Delta) = I(\exp(-x\Delta)|\Delta). \tag{B.5}$$

Taking the limit for Δ going to 0, we retrieve the Laplace transform $\mathcal{I}(x)$ corresponding to the interarrival times in the continuous-time system,

$$\mathcal{I}(x) = \lim_{\Delta \to 0} \mathcal{I}(x|\Delta) = \frac{\lambda}{x+\lambda}. \tag{B.6}$$

That is, the (batch) interarrival times are exponentially distributed. In the limit, we therefore obtain a batch Poisson arrival process characterised by the batch arrival intensity λ and by the common probability generating function of the consecutive batch sizes $\Psi(z)$.

Let $\sigma(n|\Delta)$ ($n \geq 1$) denote the probability mass function of the customer service times for the discrete-time system and given slotlength Δ. We make the discrete-time service times Δ-dependent as follows,

$$\sigma(n|\Delta) = \int_{\Delta\,(n-1)}^{\Delta\,n} s(t)dt. \tag{B.7}$$

That is, the discrete-time service times are obtained by truncating the continuous-time service times to an integer number of slots.

Let $S(z|\Delta)$ denote the probability generating function corresponding to the former probability mass function. It can then be shown that the following property holds,

$$\mathcal{S}(x) = \lim_{\Delta \to 0} S(\exp(-x\Delta)|\Delta). \tag{B.8}$$

Therefore one observes that the discrete-time distribution $\sigma(n|\Delta)$ converges to the continuous-time distribution $s(t)$ for Δ decreasing to 0. From the former expression, one also observes that the mean discretised customer service time (in absolute time) converges to the mean continuous-time customer service time $\mu_{\mathcal{S}}$, that is,

$$\lim_{\Delta \to 0} S'(1|\Delta)\,\Delta = \mu_{\mathcal{S}}. \tag{B.9}$$

Summarising, we observe that with the former definitions of $E(z|\Delta)$ and $S(z|\Delta)$, the arrival and service processes converge to a batch Poisson process and an independent service process respectively for Δ decreasing to 0.

B.3 Taking the limit

We can now plug the expressions for $E(z|\Delta)$ and $S(z|\Delta)$ into the results for the discrete-time $Geo^X/G/1$ queueing system of section 1.2. Given these arrival and

service time probability generating functions, the $Geo^X/G/1$ system reaches equilibrium whenever,

$$E'(1|\Delta)\, S'(1|\Delta) < 1. \tag{B.10}$$

Therefore, the corresponding $M^X/G/1$ system reaches equilibrium for,

$$\lim_{\Delta \to 0} E'(1|\Delta)\, S'(1|\Delta) = \lambda \mu_\Psi \mu_{\mathcal{S}} < 1. \tag{B.11}$$

Here μ_Ψ denotes the mean batch size. Equations (B.2) and (B.9) are used to obtain the former limit.

Given that the system reaches equilibrium, let $V_r(z|\Delta)$ denote the probability generating function of the queue content at random slot boundaries given slotlength Δ. From equation (1.51), we get,

$$V_r(z|\Delta) = (1 - E'(1|\Delta)\, S'(1|\Delta))\, \frac{(z-1)\, S(E(z|\Delta)|\Delta)}{z - S(E(z|\Delta)|\Delta)}. \tag{B.12}$$

The latter is also the probability generating function of the queue content at random points in time as queue content only changes just before slot boundaries. The probability generating function $\mathcal{V}_r(z)$ of the queue content at random points in time of the continuous-time system is then given by,

$$\begin{aligned} \mathcal{V}_r(z) &= \lim_{\Delta \to 0} V_r(z|\Delta) \\ &= (1 - \lambda \mu_\Psi \mu_{\mathcal{S}})\, \frac{(z-1)\, \mathcal{S}(\lambda(1 - \Psi(z)))}{z - \mathcal{S}(\lambda(1 - \Psi(z)))}. \end{aligned} \tag{B.13}$$

Here, equations (B.2), (B.8), (B.11) and (B.12) are used to obtain the limit.

Analogously, let $D(z|\Delta)$ denote the probability generating function of the customer delay for the discrete-time system, given slotlength Δ. From equation (1.36) , we retrieve,

$$D(z|\Delta) = \frac{1 - E'(1|\Delta)\, S'(1|\Delta)}{E'(1|\Delta)}\, \frac{1 - E(S(z|\Delta)|\Delta)}{z - E(S(z)|\Delta)|\Delta)}\, \frac{(z-1)\, S(z|\Delta)}{1 - S(z|\Delta)}. \tag{B.14}$$

The timing assumptions within slots imply that a customer's delay in absolute time equals the product of this customer's delay in slots and the slotlength Δ. The Laplace transform $\mathcal{D}(x)$ of the customer delay for the continuous-time system is therefore

given by,

$$\begin{aligned}\mathcal{D}(x) &= \lim_{\Delta\to 0} D\left(e^{-\Delta x}|\Delta\right) \\ &= \frac{1-\lambda\mu_\Psi\mu_{\mathcal{S}}}{\mu_\Psi}\frac{x(1-\Psi(S(x)))}{x-\lambda(1-\Psi(S(x)))}\frac{S(x)}{1-S(x)}\end{aligned} \tag{B.15}$$

Here, we used equations (B.2), (B.8) and (B.11) as well as equation (B.14) to obtain the limit.

One can verify that the former results correspond to the results for the $M^X/G/1$ queueing system, as found in literature (see a.o., Takagi [1991]). Probability generating functions of the queue content at various epochs in equilibrium as well as the Laplace transforms of unfinished work and customer waiting time can be retrieved in a similar way.

The former results show that one fairly easily retrieves performance measures for a continuous-time queueing system from performance measures of this queueing system's discrete-time counterpart.

Samenvatting

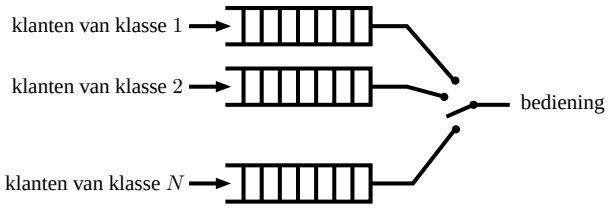
S.1 Inleiding

In een rij staan wachten is tot op zekere hoogte een deel van het dagdagelijkse leven. Of men nu aanschuift aan de kassa van een grootwarenhuis, met de wagen in de file staat of wacht tot de gevraagde webpagina op het scherm verschijnt, essentieel neemt men hetzelfde fenomeen waar. Meerdere gebruikers willen op hetzelfde ogenblik bediend worden. Het is echter zo dat slechts een beperkt aantal gebruikers tegelijkertijd bediend kan worden. Een aantal van deze gebruikers zal daarom moeten wachten. Er vormt zich dus een rij wachtende klanten: een wachtlijn.

Een abstract wachtlijnmodel beschouwt klanten (*Engels*: customers) die in een wachtlijn of buffer (*Engels*: queue, queueing system, buffer) wachten op bediening (*Engels*: service). Er kunnen één of meerdere bedieningsstations of bedieningseenheden (*Engels*: servers) zijn en ieder van de klanten vraagt een zekere bedieningstijd (*Engels*: service time) van zo een bedieningsstation. Vaak worden klanten bediend in de volgorde waarin ze in het systeem aankomen, maar het systeem kan ook uit de wachtende klanten een volgende klant kiezen volgens een ander selectiecriterium of bedieningsdiscipline (*Engels*: service discipline). Eenmaal de klant bediend is, kan deze het systeem verlaten of een andere (of eventueel dezelfde) wachtlijn vervoegen.

De wachtlijntheorie (*Engels*: queueing theory) onderzoekt wachtlijnfenomenen in een stochastisch raamwerk. Stochastische processen vatten de onzekerheid omtrent de aankomsttijdstippen van de klanten alsook de onzekerheid omtrent de bedieningstijden van deze klanten. Het is dan de taak van de wachtlijnanalist om uit de stochastische eigenschappen van deze processen prestatiematen van het wachtlijnsysteem af te leiden. Prestatiematen van belang zijn onder meer het gemiddelde en de variantie van de bufferbezetting (*Engels*: queue content) en van de wachttijden (*Engels*: customer delay) van de klanten. De bufferbezetting is gedefinieerd als het aantal klanten in het systeem, inclusief de klanten in bediening. De wachttijd van een klant is de totale tijd dat de klant in het systeem verblijft, inclusief zijn bedieningstijd.

In dit proefschrift bestuderen we een specifiek soort wachtlijnsystemen. Al de wacht-



Figuur S.1: Een generiek wachtlijnmodel met één enkel bedieningsstation en verschillende soorten klanten.

lijnsystemen die we bestuderen, beschikken over één enkel bedieningsstation dat echter slechts van tijd tot tijd beschikbaar is. In het jargon van de wachtlijntheorie spreekt men van bedieningsstations die op vakantie gaan (*Engels*: servers leave for vacations). De voornaamste wetenschappelijke bijdrage van dit proefschrift bestaat uit de prestatieanalyse van verschillende discrete-tijd-wachtlijnsystemen met vakanties. In de volgende secties bespreken we de karakteristieken van de verschillende modellen die geanalyseerd worden.

Alvorens in te gaan op deze karakteristieken, proberen we eerst dit proefschrift enigszins te situeren. In het bijzonder gaan we in op een aantal mogelijke toepassingen van wachtlijnsystemen met vakanties, lichten we de discrete-tijd-onderstelling toe en bespreken we de methode die we gebruiken bij de prestatieanalyse. Ter afsluiting van deze inleidende sectie, schetsen we tenslotte kort het onderwerp van de volgende secties.

S.1.1 Toepassingen van wachtlijnsystemen met vakanties

Wachtlijnsystemen met vakanties werden reeds veelvuldig bestudeerd in het verleden. Getuige daarvan zijn de vele referenties in zowel het overzichtsartikel van Doshi [1986] als in de boeken van Takagi [1991, 1993]. Uit de literatuur blijkt dat wachtlijnmodellen met vakanties in het bijzonder geschikt zijn voor het bepalen van de performantie van wachtlijnsystemen met verschillende soorten of klassen klanten (*Engels*: multi-class queueing systems) alsook van wachtlijnsystemen met onbetrouwbare bedieningsstations. We gaan hier dieper in op deze toepassingen.

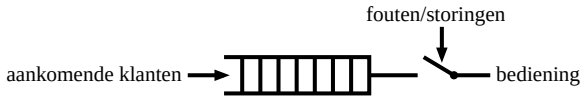
Modellen met verschillende soorten klanten

Een generiek wachtlijnsysteem met één enkel bedieningsstation en verschillende soorten of klassen klanten wordt afgebeeld in figuur S.1. De voorstelling gaat ervan uit dat klanten van verschillende klassen in verschillende wachtlijnen wachten. Klanten van een zekere klasse worden bediend overeenkomstig de stand van de schakelaar.

Klanten van alle klassen wensen bediend te worden door hetzelfde bedieningsstation. Beschouwen we nu klanten van een bepaalde klasse. Vanuit het oogpunt van deze klanten is het bedieningsstation niet beschikbaar wanneer dit station klanten van een andere klasse bedient. I.e., vanuit het oogpunt van klanten van een bepaalde klasse neemt het bedieningsstation van tijd tot tijd een vakantie. Typische voorbeelden van wachtlijnsystemen met verschillende soorten klanten zijn systemen met een prioriteitsregeling (*Engels*: priority systems, queueing systems with priorities) en polling systemen (*Engels*: polling systems).

Wachtlijnmodellen met vakanties worden onder meer aangewend door Avi-Itzhak en Naor [1963],door Nain [1983] en door Gaver Jr. [1962] om de prestatie van wachtlijnsystemen met een prioriteitsregeling te evalueren. In al deze bijdragen worden wachtlijnsystemen met preëmptieve prioriteitsdisciplines (*Engels*: preemptive priority discipline) bestudeerd. In een wachtlijnsysteem met preëmptieve prioriteiten wordt de bediening van een klant van een zekere prioriteitsklasse onmiddellijk onderbroken op het ogenblik dat er een klant van een hogere klasse aankomt. Het bedieningsstation wordt pas opnieuw beschikbaar voor klanten van de lagere klasse op het moment dat alle klanten met een hogere prioriteit bediend zijn. Na de onderbreking zijn er verschillende opties. Ofwel hervat het bedieningsstation de bediening van de klant die onderbroken werd: men noemt dit een preëmptieve prioriteitsregeling met voortzetting (*Engels*: preemptive resume priority). Een andere mogelijkheid bestaat erin dat het bedieningsstation de bediening van de klant van vooraf aan herhaalt met dezelfde bedieningstijd: men noemt deze prioriteitsregeling de preëmptieve prioriteitsregeling met herhaling (*Engels*: preemptive repeat priority). Een laatste mogelijkheid bestaat er in dat het bedieningsstation de bediening van de klant van vooraf aan herhaalt met een opnieuw bemonsterde bedieningstijd: dit is de zogenaamde preëmptieve prioriteitsregeling met herhaling en bemonstering (*Engels*: preemptive repeat with resampling priority). Zowel Avi-Itzhak en Naor [1963] als Nain [1983] gebruiken een wachtlijnmodel met vakanties om de prestatie van een wachtlijnsysteem met een preëmptieve prioriteitsregeling met voortzetting te evalueren. Gaver Jr. [1962] daarentegen beschouwt een systeem met een preëmptieve prioriteitsregeling met herhaling, en dit zowel met bemonstering als zonder. Ook in dit proefschrift worden wachtlijnsystemen met een preëmptieve prioriteitsregeling geanalyseerd (zie sectie S.2.5).

Het bedieningsstation van een polling systeem bevraagt (*Engels*: to poll) de wachtlijnen van de klanten van de verschillende klassen om beurten. Indien er klanten wachten in de bevraagde wachtlijn, bedient het station enkele van deze klanten of al deze klanten alvorens de wachtlijn van een andere klasse te bevragen. Wachtlijnmodellen met vakanties voor polling systemen worden voorgesteld en geanalyseerd door onder meer LaMaire [1991], Leung en Lucantoni [1994]en Chiarawongse et al. [1994]. Polling systemen worden door deze auteurs geanalyseerd in de context van token ring of token bus netwerken. Het bevragen komt dan overeen met het doorgeven van het token. De corresponderende vakantiemodellen zijn modellen waarbij het aantal klanten dat kan bediend worden [LaMaire, 1991] of de tijd [Leung en Lucantoni, 1994, Chiarawongse et al., 1994] tussen twee vakanties beperkt is. Wachtlijnmodellen met vakanties van



Figuur S.2: Een generiek wachtlijnmodel met één enkel onbetrouwbaar bedieningsstation.

dit type worden ook bestudeerd in sectie S.3.

Modellen met een onbetrouwbaar bedieningsstation

Wachtlijnmodellen met vakanties kunnen ook gebruikt worden om de eventuele onbetrouwbaarheid van het bedieniningsstation te modelleren. Onbetrouwbaarheid van het bedieningsstation kan resulteren in bedieningsfouten (*Engels*: service errors) of in bedieningsstoringen (*Engels*: service breakdown). Bij bedieningsfouten zal men de bediening al dan niet volledig moeten herhalen. Bij een bedieningsstoring kan de bediening van klanten slechts hervat worden nadat de uitbater van het systeem de nodige herstellingen heeft uitgevoerd. Een abstract wachtlijnsysteem met bedieningsfouten of met bedieningsstoringen wordt voorgesteld in figuur S.2. Klanten worden enkel correct bediend als de schakelaar gesloten is.

Fouten tijdens de bediening komen bijvoorbeeld voor in de context van draadloze communicatie. Transmissie van gegevens over een draadloos medium is per definitie foutgevoelig. Een draadloos communicatiesysteem kan daarom gemodelleerd worden als een bedieningsstation met vakanties. Communicatiefouten resulteren in foutief ontvangen gegevens. Door de aanwezigheid van redundantie in de verzonden berichten kan de ontvanger fouten in de transmissie opsporen en – bij foutieve transmissie – de zender vragen om de foutief ontvangen berichten opnieuw te verzenden. Hiervoor werden de zogenaamde ARQ of automatische retransmissie (*Engels*: automatic repeat request) protocollen ontwikkeld. De voornaamste ARQ protocollen zijn het stop-en-wacht (*Engels*: stop-and-wait) protocol, het ga-N-terug (*Engels*: go-back-N) protocol en het selectief-herhalen (*Engels*: selective-repeat) protocol. De prestatieanalyse van ARQ protocollen levert een aantal interessante wachtlijnproblemen op. Gezien de beperkte capaciteit van het transmissiemedium ontstaat er eerst en vooral een wachtlijn aan de kant van de zender. Verder kan het – bijvoorbeeld in het geval van het selectief-herhalen protocol – gebeuren dat de correct ontvangen berichten niet in hun oorspronkelijke volgorde aankomen bij de ontvanger. Hierdoor ontstaat er ook een wachtlijn aan de kant van de ontvanger. Reeds correct ontvangen berichten moeten wachten tot alle voorafgaande berichten correct ontvangen zijn. Towsley en Wolf [1979] en Towsley [1981] bestuderen het stop-en-wacht protocol. De prestatie van het ga-N-terug protocol wordt onderzocht door onder meer Towsley [1981] en Yoshimoto et al. [1993]. Zowel het stop-en-wacht als het ga-N-terug protocol garanderen dat de berichten in volgorde aankomen. Dit is echter niet het geval voor het selectief-herhalen protocol. Zowel Bruneel et al. [1990], Shacham en Towsley [1991],

als Kim en Krunz [2000] onderzoeken de prestatie van dit protocol. Bruneel et al. [1990] en Shacham en Towsley [1991] onderzoeken in het bijzonder de wachttijden veroorzaakt door het herordenen van berichten aan de kant van de ontvanger. Kim en Krunz [2000] daarentegen maken een benaderende analyse van de totale door de berichten ondervonden wachttijd (aan de kant van de zender en de ontvanger samen).

Onder meer Van der Duyn Schouten en Vanneste [1995], Wang et al. [1995] en Perry en Posner [2000] analyseren wachtlijnmodellen met vakanties in het kader van bedieningsstoringen. De vakanties corresponderen ofwel met de storingen ofwel met eventueel preventief onderhoud. Vaak is het zo dat de verloren tijd bij een storing veel groter is dan de nodige tijd voor preventief onderhoud. In dat geval zal het nauwkeurig plannen van het nodige preventieve onderhoud de prestatie van het systeem verbeteren.

S.1.2 Discrete-tijd-wachtlijnsystemen

Al de wachtlijnmodellen die we in dit proefschrift onderzoeken, veronderstellen dat de tijd verdeeld is in intervallen van gelijke lengte of slots. Hoewel klanten continu in het systeem kunnen aankomen, wordt er steeds van uitgegaan dat hun bediening begint op slotgrenzen. Anders gezegd, de bediening van klanten wordt gesynchroniseerd op slotgrenzen. Verder veronderstellen we ook dat de bedieningstijden steeds een geheel aantal slots bedragen zodat klanten het systeem enkel op slotgrenzen kunnen verlaten. Deze systemen worden in de literatuur gemeenzaam discrete-tijd-wachtlijnsystemen genoemd (*Engels*: discrete-time queueing systems).

De discrete tijdsschaal komt vaak overeen met de aard van de beoogde toepassing van het wachtlijnmodel. Voorbeelden zijn legio: de kloktijd in een computersysteem, de lengte van data-eenheden verzonden over een communicatiekanaal (bits, bytes, pakketten van vaste lengte), enzovoort. Verder kan men de prestatie van wachtlijnsystemen met een continue tijdsschaal bij benadering analyseren met behulp van een discrete-tijd-wachtlijnmodel door de slotlengte voldoende klein te kiezen. Alternatief kan men in veel gevallen ook op (vrij) eenvoudige wijze prestatiematen voor een continue-tijd-model afleiden uit prestatiematen voor een equivalent discrete-tijd-model door dit laatste model te beschouwen in de limiet voor slotlengtes gaande naar 0. In appendix B van dit proefschrift illustreren we deze stelling.

Discrete-tijd-wachtlijnsystemen worden reeds gedurende enkele tientallen jaren onderzocht. Vroege bijdragen zijn onder meer deze van Meisling [1958], Birdsall et al. [1962] en Powell en Avi-Itzhak [1967].Een bredere interesse voor dit specifiek soort wachtlijnproblemen ontstond slechts vanaf het einde van de jaren ’80 van de vorige eeuw. De opkomst van de asynchrone transfer mode (ATM) voor breedbandcommunicatie is hier natuurlijk niet vreemd aan.

Naslagwerken die specifiek de discrete-tijd-wachtlijntheorie behandelen zijn eerder schaars, zeker in vergelijking met het aantal beschikbare naslagwerken betreffende

continue-tijd-wachtlijntheorie (onder meer, [Takács, 1962], [Cohen, 1969], [Kleinrock, 1975, 1976] en [Cooper, 1981]). Naslagwerken over discrete-tijd-wachtlijntheorie omvatten onder meer de boeken van Bruneel en Kim [1993],Woodward [1993] en Takagi [1993]. Verder besteedt ook Hunter [1983a,b] aandacht aan een aantal discrete-tijd-wachtlijnproblemen in zijn tweedelig werk over stochastische processen met een discrete tijdsschaal. Meer recent bestudeert Daduna [2001] netwerken van wachtlijnen in discrete tijd.

S.1.3 De methode van de probabiliteitsgenererende functies

Bij de studie van de verschillende discrete-tijd-wachtlijnmodellen maken we in dit proefschrift steeds gebruik van probabiliteitsgenererende functies.

De probabiliteitsgenererende functie (*Engels*: probability generating function) of kortweg genererende functie van een (niet-negatieve) discrete toevalsgrootheid X is gedefinieerd als de z-getransformeerde van de massafunctie $x(n)$ ($n \geq 0$) van deze toevalsgrootheid,

$$X(z) = \sum_{n=0}^{\infty} x(n)\, z^n. \tag{S.1}$$

De massafunctie van de som van twee onafhankelijke discrete toevalsgrootheden is gelijk aan de convolutie van de massafuncties van deze toevalsgrootheden. Bijgevolg is de probabiliteitsgenererende functie van de som van twee onafhankelijke toevalsgrootheden gelijk aan het product van de probabiliteitsgenererende functies van deze toevalsgrootheden. Door het vermijden van convoluties, kan het gebruik van probabiliteitsgenererende functies de analyse merkelijk vereenvoudigen.

De overgang naar probabiliteitsgenererende functies impliceert echter dat de bekomen resultaten uit de analyse geen massafuncties (van toevalsgrootheden zoals bufferbezetting of wachttijden van klanten) zijn maar de overeenkomstige probabiliteitsgenererende functies. De momentgenererende eigenschap van probabiliteitsgenererende functies laat dan toe de verschillende momenten (gemiddelde waarde, variantie, ...) te bekomen door de afgeleiden van de probabiliteitsgenererende functies te evalueren voor de waarde $z = 1$. Men bekomt bijvoorbeeld de gemiddelde waarde μ_X en de variantie σ_X^2 van de toevalsgrootheid X als volgt,

$$\mu_X = X'(1), \tag{S.2}$$

$$\sigma_X^2 = X''(1) + X'(1) - X'(1)^2. \tag{S.3}$$

Verder kan men via een analyse van de singulariteiten van de probabiliteitsgenererende functie ook het staartgedrag – de massafunctie $x(n)$ voor grote waarden van n – van de overeenkomstige massafunctie benaderend bepalen.

De techniek gebaseerd op het gebruik van probabiliteitsgenererende functies is niet de enige analytische methode voor het bepalen van de performantie van discrete-tijd-wachtlijnproblemen. Een alternatief is de zogenaamde matrix-analytische methode. De analyse van een wachtlijnprobleem resulteert in dit geval in een matrixvergelijking die met behulp van efficiënte algoritmen opgelost wordt (zie onder meer Neuts [1983]).

Voor complexere systemen zal men de prestatie vaak niet meer analytisch kunnen bepalen. Men kan dan de prestatie onderzoeken met behulp van simulaties van het wachtlijnsysteem. Hoewel men de meest algemene wachtlijnproblemen kan simuleren, zal een simulatiestudie vaak veel tijd in beslag nemen. Dit is in het bijzonder het geval als men de prestatie wenst te testen voor veel verschillende parameterwaarden of als men geïnteresseerd is in gebeurtenissen die niet frequent voorkomen.

S.1.4 Overzicht

We sluiten deze inleidende sectie af met een vooruitblik op de volgende secties.

In sectie S.2 onderzoeken we wachtlijnmodellen met een willekeurig vakantieproces. De term “willekeurig” verwijst hier naar het feit dat het bedieningsstation op vakantie vertrekt, onafhankelijk van het aantal klanten in de buffer, of er al dan niet een klant in bediening is, enzovoort. Gezien zo een vakantieproces geen rekening houdt met de bediening van klanten, bestaat de mogelijkheid dat het bedieningsstation een vakantie start in het midden van de bediening van een klant. We onderzoeken in deze sectie dan ook in het bijzonder de verschillende mogelijkheden die er zijn om het onderbreken van de bediening (*Engels*: service interruption) op te vangen en de met deze verschillende mogelijkheden gepaard gaande repercussies op de prestatie van het wachtlijnsysteem. Bij wijze van toepassing tonen we aan dat wachtlijnsystemen met willekeurige vakanties gebruikt kunnen worden om de prestatie van systemen met een preëmptieve prioriteitsregeling te analyseren.

In sectue S.3 onderzoeken we wachtlijnmodellen met vakanties die niet langer onafhankelijk van de toestand van het systeem genomen worden. Het eerste model dat we onderzoeken is een vakantiemodel met een poort en met een exhaustieve bediening (*Engels*: gated-exhaustive vacation system) . Dit model combineert eigenschappen van zowel het vakantiesysteem voorzien van een poort (*Engels*: gated vacation system) als het vakantiesysteem met exhaustieve bediening (*Engels*: exhaustive vacation system). In het vakantiesysteem met een poort vertrekt het bedieningsstation op vakantie als alle klanten bediend zijn die aanwezig waren in het systeem bij de terugkeer van een vorige vakantie. Het bedieningsstation van het vakantiesysteem met exhaustieve bediening vertrekt op vakantie als er geen klanten meer in het systeem wachten. Daarnaast bestuderen we ook een systeem met een vrij algemeen vakantieproces – er is onder meer correlatie mogelijk tussen opeenvolgende vakanties en de bediening van klanten kan onderbroken worden – en tonen aan dat we met dit model de prestatie kunnen analyseren van een heel aantal “klassieke” vakantiemodellen zonder poorten.

S.2 Willekeurige vakanties

In deze sectie bestuderen we wachtlijnmodellen met een willekeurig vakantieproces. De term *willekeurig* slaat hier op het feit dat het al dan niet nemen van vakanties niet afhangt van de staat van het wachtlijnsysteem. I.e., het bedieningsstation vertrekt op vakantie of er nu al dan niet klanten aan het wachten zijn op bediening, of er nu al dan niet een klant bediend wordt, enzovoort. Gezien de bediening van klanten onderbroken kan worden, spreekt men ook wel eens van onderbrekingsmodellen (*Engels*: interruption models) of van wachtlijnsystemen met bedieningsonderbrekingen (*Engels*: queueing systems with service interruptions).

Discrete-tijd-wachtlijnmodellen met willekeurige vakanties werden reeds veelvuldig geanalyseerd in het verleden. Van eenvoudige Bernoulli-onderbrekingsmodellen (bijvoorbeeld [Hsu, 1974] en [Heines, 1979]) tot modellen met correlatie in het onderbrekingsproces (onder meer [Bruneel, 1986] en Yang en Mark [1990]) en modellen met meerdere bedieningsstations (bijvoorbeeld [Georganas, 1976] en Laevens en Bruneel [1995]).Al deze bijdragen gaan er echter van uit dat de bediening van een klant slechts één enkel slot in beslag neemt. Inghelbrecht et al. [2000] bestuderen een wachtlijnmodel met onderbrekingen en bedieningstijden met een vaste lengte. Wachtlijnmodellen met onderbrekingen en willekeurige bedieningstijden worden bestudeerd in deze sectie.

Deze sectie vat hoofdstuk 2 van dit proefschrift samen en is gebaseerd op de volgende publicaties: [Fiems et al., 2001], [Fiems et al., 2002b], [Fiems et al., 2002c] en [Fiems et al., 2003].

S.2.1 Algemene onderstellingen

We gaan nu eerst in op een aantal veronderstellingen die gelden voor alle wachtlijnmodellen die we in deze sectie analyseren.

Zoals reeds aangegeven in de inleiding beschouwen we in dit proefschrift discrete-tijd-wachtlijnmodellen. We veronderstellen dat de tijd onderverdeeld is in intervallen van gelijke lengte, ook wel slots genoemd. Klanten vervoegen het wachtlijnsysteem gedurende de opeenvolgende slots en de wachtlijn biedt plaats aan een onbeperkt aantal klanten. De klanten worden volgens een eerst-in-eerst-uit-regeling (*Engels*: first-in-first-out of FIFO scheduling discipline) bediend. I.e., klanten die op een vroeger tijdstip aankomen, worden eerst bediend.

Het aantal klanten dat gedurende de opeenvolgende slots in het wachtlijnsysteem aankomt, wordt gemodelleerd met behulp van een reeks van onafhankelijke identiek gedistribueerde (*Engels*: independent identically distributed of i.i.d.) niet-negatieve toevalsgrootheden. Het aankomstproces wordt daarom volledig gekarakteriseerd door de

massafunctie van het aantal aankomsten in een slot of, alternatief, door de probabiliteitsgenererende functie die met deze massafunctie correspondeert.

Bediening van klanten wordt gesynchroniseerd op slotgrenzen. Bijgevolg kan een klant niet bediend worden gedurende zijn aankomstslot. We veronderstellen verder dat de bedieningstijden van de opeenvolgende klanten een reeks onafhankelijke identiek gedistribueerde positieve toevalsgrootheden zijn, gekarakteriseerd door de gemeenschappelijke massafunctie of door de overeenkomstige probabiliteitsgenererende functie.

Er is één enkel bedieningsstation. Dit is echter niet steeds beschikbaar. De stochastische processen die de verschillende onderzochte vakantieprocessen beschrijven, worden toegelicht in de volgende sectie. Het is mogelijk dat het bedieningsstation op vakantie vertrekt tijdens het bedienen van een klant. We beschouwen daarom een aantal werkingsmodi (*Engels*: operation mode) die verschillen in de manier waarop de klant zijn bediening na een onderbreking hervat. De verschillende werkingsmodi die in dit proefschrift onderzocht worden, worden beschreven in sectie S.2.3.

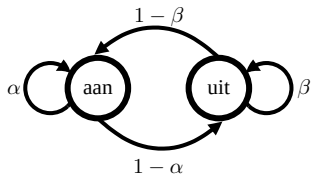
S.2.2 Vakantiemodellen

We bestuderen in deze sectie de volgende drie vakantiemodellen: het Bernoulli-model, het Markoviaanse model en het aan/uit-model met algemeen gedistribueerde uit-tijden. De stochastische karakteristieken van deze verschillende modellen worden in deze sectie besproken.

Het Bernoulli-model

Het Bernoulli-vakantiemodel veronderstelt dat de kans dat het bedieningsstation op vakantie is tijdens een slot onafhankelijk is van het eventueel op vakantie zijn van het station gedurende voorafgaande slots. Aangezien we systemen met één enkel bedieningsstation beschouwen, zal het aantal beschikbare bedieningsstations gedurende de opeenvolgende slots een reeks onafhankelijke identiek Bernoulli-gedistribueerde toevalsgrootheden vormen. Dit vakantiemodel wordt daarom volledig gekarakteriseerd door de kans σ dat het bedieningsstation beschikbaar is gedurende een willekeurig slot.

Het vakantieproces kan alternatief beschreven worden als een aan/uit-proces met onafhankelijke (verschoven) geometrisch verdeelde aan- en uit-tijden. Het bedieningsstation is beschikbaar gedurende aan-tijden en op vakantie gedurende uit-tijden. De massafuncties $a(n)$ en $b(n)$ ($n > 0$) van respectievelijk de aan- en de uit-tijden worden



Figuur S.3: Transitiediagram van het Markoviaanse vakantieproces.

gegeven door

$$a(n) = \sigma^{n-1}(1-\sigma), \tag{S.4}$$
$$b(n) = (1-\sigma)^{n-1}\sigma. \tag{S.5}$$

Aangezien het onderbrekingsproces volledig gekarakteriseerd is door één enkele parameter, kunnen we de gemiddelde aan-tijd en de gemiddelde uit-tijd niet onafhankelijk van elkaar kiezen.

Het Markoviaanse vakantiemodel

In het geval van Bernoulli-onderbrekingen zijn zowel de aan- als de uit-tijden (verschoven) geometrisch gedistribueerd. Verder is het niet mogelijk de gemiddelde lengte van aan- en uit-periodes onafhankelijk te kiezen. Het Markoviaanse vakantiemodel gaat nog steeds uit van (verschoven) geometrisch gedistribueerde aan- en uit-tijden. We kunnen echter de gemiddelde lengte van de aan- en uit-tijden wel onafhankelijk kiezen. De massafuncties $a(n)$ en $b(n)$ ($n > 0$) van respectievelijk de aan- en de uit-tijden worden nu gegeven door

$$a(n) = \alpha^{n-1}(1-\alpha), \tag{S.6}$$
$$b(n) = \beta^{n-1}(1-\beta), \tag{S.7}$$

waarbij α en β willekeurige kunnen gekozen worden tussen 0 en 1.

De (verschoven) geometrische verdeling heeft geen geheugen. Dit wil onder meer zeggen dat de kans dat de aan- of uit-tijd verder gaat gedurende het volgende slot onafhankelijk is van het aantal slots dat de aan- of uit-tijden reeds duren. Gegeven dat het bedieningsstation beschikbaar is gedurende een slot, zal het bedieningsstation beschikbaar blijven gedurende het daaropvolgende slot met kans α. Analoog, blijft het bedieningsstation op vakantie met kans β gegeven dat het station op vakantie was. Bijgevolg kan men het Markoviaanse vakantiemodel karakteriseren aan de hand van een transitiediagram zoals getoond in figuur S.3.

In tegenstelling tot het Bernoulli-model is de beschikbaarheid van het bedieningsstation gedurende een bepaald slot niet langer onafhankelijk van het al dan niet beschikbaar zijn van dit station gedurende het voorafgaande slot. Het vakantieproces is gecorreleerd in de tijd.

Het aan/uit-model met algemeen gedistribueerde uit-tijden

Het laatste vakantiemodel dat we in deze sectie bestuderen is een aan/uit-model met (verschoven) geometrisch verdeelde aan-tijden en algemeen verdeelde uit-tijden. Aangezien de geometrische verdeling geen geheugen heeft, is de kans dat het bedieningsstation na een aan-slot op vakantie vertrekt constant en dus onafhankelijk van de duur van de aan-tijd op dat tijdstip. Het vakantieproces wordt nu volledig gekarakteriseerd door de kans α dat het bedieningsstation beschikbaar blijft en de massafunctie van de uit-tijden.

Het al dan niet beschikbaar zijn van het bedieningsstation gedurende een bepaald slot bepaalt in het geval van Markoviaanse onderbrekingen de kans dat het bedieningsstation beschikbaar is in het daaropvolgende slot. Dit is niet langer het geval voor dit algemenere aan/uit-proces.

S.2.3 Onderbroken bediening

Zoals reeds vermeld werd, impliceert de combinatie van willekeurige bedieningstijden en vakanties dat het bedieningsstation op vakantie kan vertrekken op een moment dat er een klant bediend wordt. De bediening van de klant wordt dan onderbroken. Na de vakantie kan het bedieningsstation de bediening van de onderbroken klant op verschillende manieren hervatten. De verschillende bestudeerde mogelijkheden – de werkingsmodi – worden in deze sectie besproken.

Drie modi

In de eerste werkingsmodus zet het bedieningsstation de bediening verder van de klant wiens bediening onderbroken werd. De bediening die de klant reeds kreeg voor de onderbreking gaat in dit geval niet verloren. Deze werkingsmodus wordt de ga-verder-na-de-onderbreking (*Engels*: Continue After Interruption of CAI) werkingsmodus genoemd.

In een tweede werkingsmodus herhaalt het bedieningsstation de bediening van de klant wiens bediening onderbroken werd van vooraf aan. In deze werkingsmodus gaat de door de klant ontvangen bediening voorafgaand aan de onderbreking verloren. Deze werkingsmodus wordt de herhaal-na-de-onderbreking (*Engels*: Repeat After Interruption of RAI) werkingsmodus genoemd.

In de derde werkingsmodus tenslotte, wordt net zoals in de tweede modus de bediening van de klant van vooraf aan herhaald. De bedieningstijd van deze klant wordt echter na iedere onderbreking opnieuw bemonsterd. Dit wil zeggen dat de bedieningstijd na de onderbreking een toevalsgrootheid is met dezelfde distributie als en onafhankelijk van de oorspronkelijke bedieningstijd. Deze werkingsmodus wordt de herhaal-en-bemonster-na-de-onderbreking (*Engels*: Repeat After Interruption with resampling of RAI,wr) genoemd.

In dit proefschrift combineren we deze 3 werkingsmodi met de voorafgaande 3 vakantiemodellen.

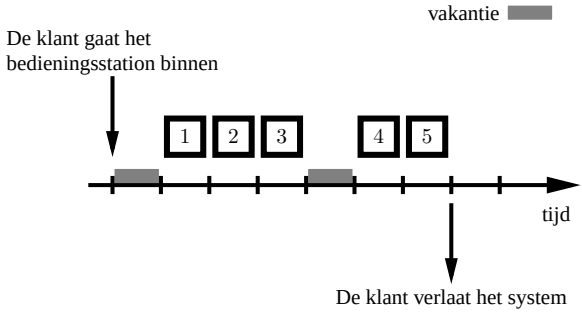
Enkele varianten

Naast de voorafgaande werkingsmodi, onderzoeken we ook een aantal varianten. In tegenstelling tot de voorafgaande modi, worden deze varianten enkel onderzocht in het geval van een Bernoulli-vakantiemodel.

Bij vertraagde werkingsmodi worden klanten verder bediend tijdens vakantieperiodes. De klant verlaat echter enkel het bedieningsstation indien dit station beschikbaar was gedurende de volledige bedieningstijd van deze klant. I.e., de bediening van de klant wordt herhaald totdat het bedieningsstation gedurende de hele bedieningstijd beschikbaar is. We beschouwen zowel het geval waarin de bedieningstijd dezelfde blijft bij (eventuele) herhalingen en het geval waar deze bedieningstijd steeds opnieuw bemonsterd wordt. We refereren naar deze werkingsmodi als de d-RAI (*Engels*: delayed Repeat After Interruption) en de d-RAI,wr (*Engels*: delayed Repeat After Interruption with resampling) werkingsmodi.

Voor de modi RAI, RAI,wr, d-RAI en d-RAI,wr wordt bij onderbrekingen steeds de volledige bedieningstijd (eventueel opnieuw bemonsterd) herhaald. In het geval van partiële werkingsmodi wordt er van uit gegaan dat de bedieningstijd opgesplitst is in delen en dat eenmaal een deel van de bediening is afgewerkt, het niet meer herhaald wordt. Anders gezegd, we passen de verschillende werkingsmodi toe op de delen van de bedieningstijd. De partiële modi die overeenkomen met RAI, RAI,wr, d-RAI en d-RAI,wr worden aangeduid met p-RAI (*Engels*: partial RAI), p-RAI,wr (*Engels*: partial RAI,wr), dp-RAI (*Engels*: delayed partial RAI) en dp-RAI,wr (*Engels*: delayed partial RAI,wr).

Voor deze partiële modi dienen we verdere onderstellingen te maken betreffende de manier waarop de bedieningstijd van een klant in delen wordt opgesplitst. We onderstellen in dit proefschrift dat het aantal delen in de opeenvolgende bedieningstijden en de lengte (in slots) van deze delen reeksen onafhankelijke identiek gedistribueerde positieve toevalsgrootheden zijn.



Figuur S.4: Effectieve bedieningstijd voor de CAI modus.

S.2.4 De methode van de effectieve bedieningstijden

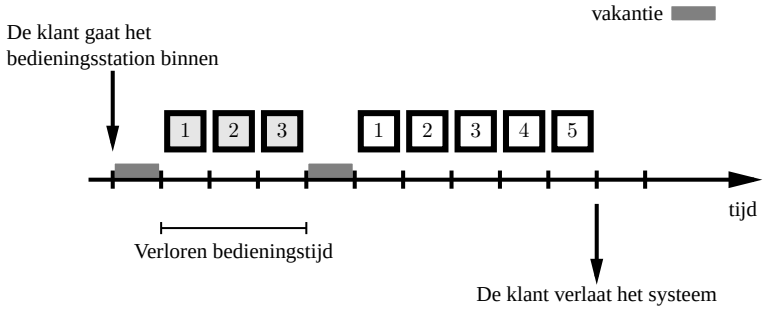
De verschillende voorgestelde wachtlijnmodellen worden geanalyseerd met de methode van de effectieve bedieningstijden. Deze methode bestaat uit twee stappen. In een eerste stap bepalen we de stochastische karakteristieken van de zogenaamde effectieve bedieningstijden. In een tweede stap wordt dan de verdere wachtlijnanalyse uitgevoerd met behulp van de bekomen karakteristieken. De verdere wachtlijnanalyse is echter een stuk eenvoudiger door het gebruik van deze effectieve bedieningstijden. In deze sectie lichten we de methode van de effectieve bedieningstijden toe en vergelijken deze methode met de methode van de toegevoegde toestandsvariabelen.

Effectieve bedieningstijden

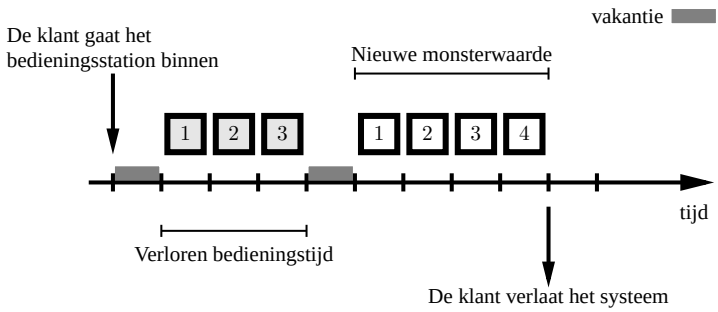
De effectieve bedieningstijd van een klant is gedefinieerd als het aantal slots tussen het begin van het slot waar de klant het bedieningsstation binnengaat en het einde van het slot waar de klant het bedieningsstation (en dus ook het systeem) verlaat. Een klant gaat het bedieningsstation binnen bij het begin van het slot volgend op het vertrekslot van de voorafgaande klant als er bij zijn aankomst andere klanten in het systeem aanwezig zijn of bij het begin van het slot volgend op zijn aankomstslot indien dit niet het geval is. Merk op dat het binnengaan van het bedieningsstation niet impliceert dat de klant ook onmiddellijk bediend wordt. Het is immers mogelijk dat het bedieningsstation op vakantie is op het moment dat de klant het station binnengaat.

Figuren S.4, S.5 en S.6 illustreren deze definitie voor de CAI, RAI en RAI,wr werkingsmodi respectievelijk. We veronderstellen dat de bedieningstijd van de klant 5 slots bedraagt. In het geval van RAI,wr, wordt deze bedieningstijd opnieuw bemonsterd tot 4 slots. Merk op dat in dit specifiek geval het bedieningsstation niet beschikbaar is op het ogenblik dat de klant het bedieningsstation binnengaat.

In dit proefschrift bepalen we eerst de probabiliteitsgenererende functies van de effec-



Figuur S.5: Effectieve bedieningstijd voor de RAI modus.



Figuur S.6: Effectieve bedieningstijd voor de RAI,wr modus.

tieve bedieningstijden voor de verschillende werkingsmodi. Voor het Bernoulli-model ziet men vrij eenvoudig in dat de effectieve bedieningstijden van de opeenvolgende klanten een reeks onafhankelijke identiek gedistribueerde toevalsgrootheden vormen. Dit is een gevolg van de afwezigheid van correlatie in het vakantieproces en van de afwezigheid van correlatie tussen de bedieningstijden van de opeenvolgende klanten. Voor het Markoviaanse vakantiemodel en het vakantiemodel met algemeen gedistribueerde vakanties is dit niet meer het geval. Er is immers correlatie in het vakantieproces.

Voor het Markoviaanse vakantiemodel beschrijft het al dan niet beschikbaar zijn van het bedieningsstation gedurende een slot de toestand (in de Markoviaanse betekenis) van het vakantieproces volledig. De effectieve bedieningstijd van een klant is een toevalsgrootheid die afhangt van de toestand van het vakantieproces gedurende het slot voorafgaand aan deze effectieve bedieningstijd. Daarom bepalen we voor het Markoviaanse model de (conditionele) probabiliteitsgenererende functies van de effectieve bedieningstijden gegeven dat het bedieningsstation beschikbaar is gedurende het slot dat aan de effectieve bedieningstijd voorafgaat en gegeven dat dit niet het geval is. We leiden verder ook een verband af tussen deze genererende functies af dat onafhankelijk is van de werkingsmodus.

De beschikbaarheid van het bedieningsstation gedurende een slot beschrijft de toestand van het vakantieproces slechts gedeeltelijk voor het model met algemene vakantietijden. Een volledige toestandsbeschrijving bestaat bijvoorbeeld uit het al dan niet beschikbaar zijn van het bedieningsstation in combinatie met het aantal slots dat het station nog op vakantie zal blijven indien het station niet beschikbaar is. Opnieuw kunnen we de (conditionele) probabiliteitsgenererende functies van de effectieve bedieningstijden bepalen gegeven de toestand gedurende het slot voorafgaand aan de effectieve bedieningstijd. Voor de verdere analyse kunnen we ons in dit geval beperken tot het bepalen van de probabiliteitsgenererende functie van de effectieve bedieningstijden gegeven dat het bedieningsstation beschikbaar is gedurende het slot voorafgaand aan de effectieve bedieningstijd (zie verder).

De probabiliteitsgenererende functies van de effectieve bedieningstijden voor de verschillende werkingsmodi en vakantieprocessen kunnen we – kort samengevat – bepalen uit een set van recursieve vergelijkingen voor de massafuncties van de effectieve bedieningstijden, geconditioneerd op de bedieningstijden van de klanten. Deze recursieve vergelijkingen bekomen we door te conditioneren op de toestand van het vakantieproces gedurende het eerste slot van de effectieve bedieningstijden. De probabiliteitsgenererende functies van de effectieve bedieningstijden worden op deze wijze bepaald voor het Markoviaanse vakantieproces en voor het proces met algemeen gedistribueerde vakanties. Voor het Bernoulli-vakantieproces gebruiken we in dit proefschrift – gezien de eenvoud van het Bernoulli-vakantieproces – een ad-hoc methode om de genererende functie van de effectieve bedieningstijd te bepalen.

Een wachtlijnprobleem zonder vakanties

Eenmaal we de stochastische karakteristieken (de probabiliteitsgenererende functies) van de effectieve bedieningstijden achterhaald hebben, reduceert het wachtlijnprobleem met vakanties zich tot een wachtlijnprobleem zonder vakanties maar met bedieningstijden gegeven door de effectieve bedieningstijden. Het gereduceerde wachtlijnprobleem is vaak eenvoudiger dan het originele probleem. Los daarvan hangt het gereduceerde probleem niet af van de beschouwde werkingsmodus. De verdere wachtlijnanalyse kan dus simultaan voor de verschillende werkingsmodi uitgevoerd worden.

In het geval van Bernoulli-onderbrekingen vormen de opeenvolgende effectieve bedieningstijden een reeks onafhankelijke identiek gedistribueerde toevalsgrootheden. Het wachtlijnprobleem reduceert zich in dit geval tot een standaard wachtlijnprobleem met onafhankelijk aankomstproces en algemene onafhankelijke bedieningstijden. Zo een systeem werd onder meer bestudeerd door Bruneel [1993] en Takagi [1993] alsook in het inleidende hoofdstuk van dit proefschrift.

Voor het Markoviaanse vakantieproces is dit niet het geval aangezien de effectieve bedieningstijden afhangen van de toestand waarin het vakantieproces zich bij het begin van de effectieve bedieningstijd bevindt. We kunnen echter opmerken dat het

bedieningsstation per definitie beschikbaar is gedurende het laatste slot van de effectieve bedieningstijd. Bijgevolg kunnen effectieve bedieningstijden enkel voorafgegaan worden door een slot waar het bedieningsstation op vakantie is als de klant aankomt in een leeg systeem. Voor dit wachtlijnsysteem bepalen we onder meer de probabiliteitsgenererende functies van de bufferbezetting op vertrektijdstippen en op willekeurige slotgrenzen in stochastisch evenwicht alsook de probabiliteitsgenererende functie van de wachttijd van een willekeurige klant in stochastisch evenwicht.

Voor het aan/uit-vakantiemodel met algemeen gedistribueerde vakantietijden kunnen we dezelfde methode volgen als in het geval van Markoviaanse onderbrekingen. In dit proefschrift volgen we echter een alternatieve methode. We tonen eerst aan dat de genererende functies van de effectieve bedieningstijden gegeven dat het bedieningsstation beschikbaar is gedurende het slot voorafgaand aan de effectieve bedieningstijd gelijk is aan de probabiliteitsgenererende functie van de uitgebreide afwerkingstijd (*Engels*: extended service completion time) van een klant. De afwerkingstijd (*Engels*: service completion time) van een klant is de tijd (in slots) tussen het begin van het slot waar de klant voor het eerst bediend wordt en het einde van het slot waar de klant het systeem verlaat. De uitgebreide afwerkingstijd is gelijk aan de afwerkingstijd in het geval het bedieningsstation beschikbaar is gedurende het slot na de afwerkingstijd en is gelijk aan de som van de afwerkingstijd en de daaropvolgende vakantie indien dit niet het geval is. We kunnen dan opmerken dat het exhaustief vakantiemodel met meerdere vakanties (zie sectie S.3), met bedieningstijden gelijk aan de uitgebreide afwerkingstijden en met een aangepaste vakantiedistributie bijna equivalent functioneert als het originele systeem. Een vergelijking van beide systemen laat dan toe de probabiliteitsgenererende functies van de bufferbezetting en van de wachttijden te bepalen. Verder bepalen we voor dit vakantiemodel ook op eenvoudige wijze de genererende functie van de periodes waar het bedieningsstation noch klanten bedient noch op vakantie is, i.e., de werkloze periodes (*Engels*: idle period). Ook bepalen we de genererende functie van de periodes waar het bedieningsstation of een klant bedient of op vakantie is, i.e., de bezige periodes (*Engels*: busy period). Gezien de karakteristieken van het aankomstproces, van de bedieningstijden en van het vakantieproces, vormen de opeenvolgende werkloze en bezige periodes twee onafhankelijke rijen onafhankelijke toevalsgrootheden. De werkloze periodes zijn bovendien geometrisch verdeeld.

Een alternatieve oplossingsmethode

De methode van de effectieve bedieningstijden is natuurlijk niet de enige methode om wachtlijnsystemen met vakanties te analyseren. Een veelgebruikte methode in de wachtlijntheorie is de methode van de toegevoegde variabelen (*Engels*: method of the supplementary variables). In het algemeen zal, gegeven de bufferbezetting bij het begin van een slot, het toekomstig gedrag van het wachtlijnsysteem afhangen van het voorbije. Vaak kunnen we echter naast de bufferbezetting een beperkt aantal toestandsvariabelen (*Engels*: state variables) bijhouden. Gegeven de bufferbezetting en

deze toestandsvariabelen wordt het toekomstig gedrag onafhankelijk van het voorafgaand gedrag van het wachtlijnsysteem. I.e., de bufferbezetting en de toestandsvariabelen vormen een toestandsvector in de Markoviaanse betekenis. Eenmaal men de stochastische karakteristieken van deze toestandsvector in stochastisch evenwicht achterhaald heeft, kan men op eenvoudige wijze een aantal prestatiematen van het wachtlijnsysteem afleiden. Zowel Stidham [2002] als Takagi [1991] schrijven deze methode toe aan Cox [1955].

We passen de methode van de toegevoegde variabelen toe in het geval van Bernoulli-vakanties en voor de CAI, de RAI en de RAI,wr werkingsmodus.

In vergelijking met de methode van de toegevoegde variabelen is de methode van de effectieve bedieningstijden compacter. We kunnen immers een groot gedeelte van de analyse simultaan uitvoeren voor al de verschillende onderzochte werkingsmodi. De methode van de effectieve bedieningstijden is echter een ad-hoc methode, specifiek gericht op de analyse van vakantiesystemen. Het aantal wachtlijnproblemen dat via de methode van de toegevoegde variabelen kan opgelost worden is veel groter. Deze methode vertaalt het gestelde wachtlijnprobleem op een redelijk eenvoudige wijze in een vergelijking voor de gezamenlijke genererende functie van de toestand van het systeem. Eventueel onbepaalde constanten en/of functies in deze vergelijkingen tracht men te bepalen met behulp van de normalisatievoorwaarde voor genererende functies en het feit dat (partiële) genererende functies analytisch zijn binnen de eenheidscirkel. Hoewel de te manipuleren formules vaak lang zijn, brengt dit op zich weinig problemen mee. De beschikbaarheid van algebraïsche wiskundige computerapplicaties is hier niet vreemd aan.

S.2.5 Toepassing: een preëmptief prioriteitsmodel

Wachtlijnmodellen met vakanties kunnen gebruikt worden om de prestatie te schatten van wachtlijnen met een preëmptieve prioriteitsregeling.

Wachtlijnmodel

We onderzoeken een wachtlijn met N verschillende soorten – N verschillende klassen – klanten. We veronderstellen dat klanten van klasse 1 de hoogste prioriteit hebben en klanten van klasse N de laagste. We beschouwen een preëmptief prioriteitssysteem. Preëmptie wil zeggen dat de bediening van klanten van een zekere klasse onderbroken wordt indien er klanten met een hogere prioriteit in het systeem aankomen. Pas als alle klanten met hogere prioriteit bediend zijn, wordt het bedieningsstation opnieuw beschikbaar voor klanten met lagere prioriteit. Zoals reeds in de inleidende sectie vermeld is, kan de klant na een onderbreking op verschillende manieren de bediening hernemen. Ofwel zet het bedieningsstation de bediening verder van de klant die onderbroken werd: dit is preëmptieve prioriteitsregeling met voortzetting. Ofwel herhaalt

het bedieningsstation de bediening van de klant van vooraf aan met de zelfde bedieningstijd: dit is de preëmptieve prioriteitsregeling met herhaling. Ofwel herhaalt het bedieningsstation de bediening van de klant van vooraf aan met een opnieuw bemonsterde bedieningstijd: dit is de preëmptieve prioriteit met herhaling en bemonstering.

We veronderstellen verder dat de aantallen aankomsten van een bepaalde klasse gedurende de opeenvolgende slots een reeks onafhankelijke identiek gedistribueerde niet-negatieve toevalsgrootheden vormen. De reeksen voor de verschillende klassen zijn ook onderling onafhankelijk. De bedieningstijden van de klanten van een bepaalde klasse vormen een reeks onafhankelijke identiek gedistribueerde positieve toevalsgrootheden. Opnieuw zijn deze reeksen voor de verschillende klassen ook onderling onafhankelijk. Merk op dat we voor elke klasse verschillende distributies kunnen kiezen voor het aantal aankomsten van die klasse in een slot. Analoog kunnen we ook verschillende distributies voor de bedieningstijden van de klanten van de verschillende klassen kiezen.

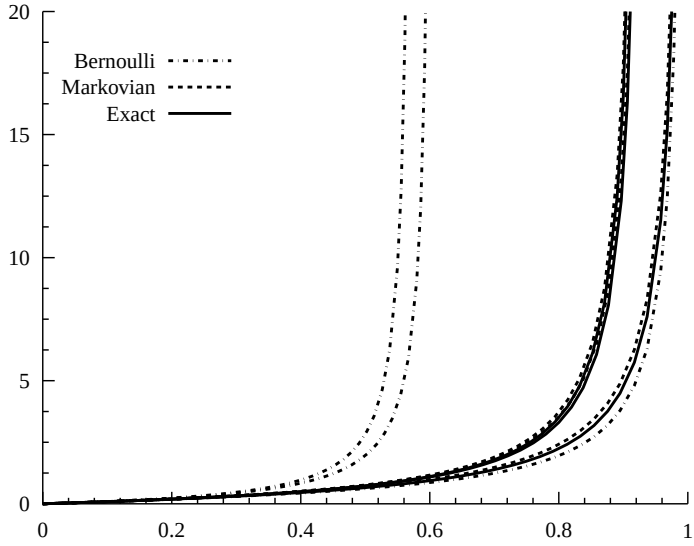
Exacte analyse

Het model met algemeen verdeelde vakanties laat toe systemen met een preëmptieve prioriteitsregeling exact te modelleren.

Beschouw eerst klanten van klasse 1. De preëmptieve prioriteitsregeling impliceert dat klanten van klasse 1 bediend worden alsof er geen andere klanten in het systeem aanwezig zijn. Bij gevolg kan men de performantie voor klanten van klasse 1 bepalen met behulp van een model zonder prioriteiten of vakanties.

Klanten van klasse 2 kunnen daarentegen slechts bediend worden als er geen klanten van klasse 1 in het systeem aanwezig zijn. Voor klanten van klasse 2 is het bedieningsstation afwisselend beschikbaar en op vakantie. De periodes waar het bedieningsstation beschikbaar is, komen overeen met periodes waar er geen klanten van klasse 1 in het systeem zijn, i.e., met de werkloze periodes van klasse 1. De periodes waar het bedieningsstation op vakantie is komen overeen met de periodes waar klanten van klasse 1 in het systeem aanwezig zijn, i.e., met de bezige periodes van klasse 1. De opeenvolgende werkloze en bezige periodes van de klanten van klasse 1 vormen reeksen onafhankelijke toevalsgrootheden. De werkloze periodes zijn bovendien geometrisch verdeeld. We kunnen dus de performantie voor klanten van klasse 2 bepalen met behulp van het aan/uit-vakantiemodel met geometrisch verdeelde beschikbare periodes en algemeen verdeelde vakanties. Afhankelijk van de beschouwde werkingsmodus bekomen we de verschillende preëmptieve prioriteitsregelingen. De CAI modus komt overeen met de preëmptieve prioriteitsregeling met voortzetting, de RAI modus komt overeen met de preëmptieve prioriteitsregeling met herhaling en de RAI,wr modus komt overeen met de preëmptieve prioriteitsregeling met herhaling en bemonstering.

Voor lagere prioriteitsklassen kunnen we analoog te werk gaan. Hierbij maken we gebruik van het feit dat de opeenvolgende bezige en werkloze periodes van alle hogere



Figuur S.7: Gemiddelde bufferbezetting van klanten van klasse 2 in functie van de totale systeembelasting. (2 prioriteitsklassen; klasse 1: Bernoulli-aankomsten, verschoven symmetrisch binomiaal verdeelde bedieningstijden met gemiddelde $\mu_S = 10$, 20% van de belasting; klasse 2: Bernoulli-aankomsten, verschoven symmetrisch binomiaal verdeelde bedieningstijden met gemiddelde $\mu_S = 10$; 80% van de belasting.)

prioriteitsklassen samen een reeks onafhankelijke en een reeks geometrisch verdeelde onafhankelijke toevalsgrootheden vormen. Merk hierbij op dat de bezige periodes van een bepaalde klasse ook de vakanties die ervaren worden door die klasse omvatten.

Bijgevolg kunnen we met behulp van het vakantiemodel met algemeen verdeelde vakantietijden de prestatie bepalen van een wachtlijnsysteem met een preëmptieve prioriteitsregeling en een willekeurig aantal klassen. We kunnen voor elke klasse kiezen of we na een onderbreking de bediening voortzetten, herhalen of herhalen en opnieuw bemonsteren. Het is bijvoorbeeld mogelijk dat klanten van klasse 2 na een onderbreking hun bediening verderzetten terwijl klanten van klasse 3 hun bediening herhalen na een onderbreking.

Benaderende analyse

We gaan ook na in welke mate het Bernoulli- en het Markoviaanse model kunnen gebruikt worden om dit soort prioriteitssystemen benaderend te modelleren. We beperken ons tot een systeem met 2 prioriteitsklassen.

Voor de benadering met het Bernoulli-vakantiemodel stellen we de kans σ dat het bedieningsstation beschikbaar is gedurende een willekeurig slot, gelijk aan de kans dat de buffer van de hogere prioriteitsklasse leeg is gedurende een willekeurig slot. Voor

het Markoviaanse model stellen we de gemiddelde aan-tijd en de gemiddelde vakantietijd gelijk aan respectievelijk de gemiddelde tijd dat de hogere prioriteitsbuffer leeg is en de gemiddelde tijd dat dit niet het geval is.

In figuur S.7 zetten we de gemiddelde bufferbezetting van klanten van klasse 2 uit in functie van de systeembelasting voor – van links naar rechts – de preëmptieve prioriteitsregeling met herhaling, de preëmptieve prioriteitsregeling met herhaling en bemonstering en de preëmptieve prioriteitsregeling met voortzetting. Voor de preëmptieve prioriteitsregeling met voortzetting zijn beide benaderingen vrij goed. Voor de andere regelingen levert enkel het Markoviaanse model een goede benadering. Voor de variantie van de bufferbezetting (zie hoofdstuk 2) blijkt het Markoviaanse model nog steeds de exacte waarde vrij goed te benaderen. Het Bernoulli-model daarentegen levert voor geen enkel van de drie prioriteitsregelingen een goede benadering.

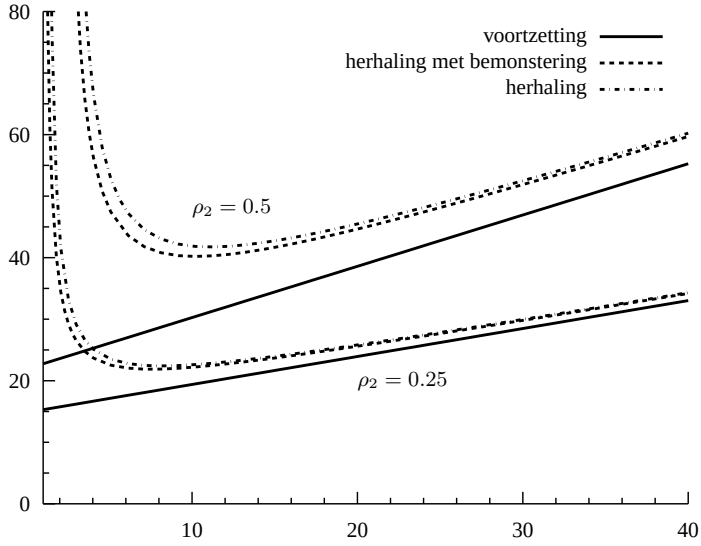
De lengte van de bedieningstijden

Erg opvallend is de invloed van de gemiddelde lengte van de bedieningstijden van klanten van klasse 1 (bij gelijkblijvende belasting) op de performantie van de klanten van klasse 2 in een wachtlijnsysteem met 2 prioriteitsklassen en preëmptieve prioriteiten.

In figuren S.8 en S.9 worden respectievelijk het gemiddelde en de variantie van de wachttijden van klasse 2 uitgezet in functie van de gemiddelde lengte van de bedieningstijden van klanten van klasse 1. We onderstellen een Bernoulli-aankomstproces voor zowel klanten van klasse 1 als van klasse 2. De bedieningstijden van klasse 1 zijn geometrisch verdeeld. De bedieningstijden van klanten van klasse 2 daarentegen zijn verschoven (symmetrisch) binomiaal verdeeld. De gemiddelde bedieningstijd van een klant van klasse 2 bedraagt 10 slots. De klasse-1-belasting bedraagt 20% en we beschouwen verschillende waarden voor de klasse-2-belasting ρ_2.

Voor de preëmptieve prioriteitsregeling met voortzetting ziet men dat zowel de gemiddelde waarde als de variantie van de wachttijden van klanten van klasse 2 stijgen voor langere bedieningstijden van klanten van klasse 1. Dit is een gevolg van het feit dat de werkperiodes van klasse 1 (i.e., de vakantieperiodes voor klasse 2) langer worden voor langere bedieningstijden van klanten van klasse 1. Met andere woorden, het onderbrekingsproces dat door klanten van klasse 2 ondervonden wordt, wordt grilliger (langere aan- en uit-tijden) als de gemiddelde bedieningstijd van klanten van klasse 1 toeneemt. De wachtlijn van de klanten van klasse 2 groeit verder aan gedurende deze langere uit-periodes waardoor de performantie van het systeem vanuit het perspectief van de klanten van klasse 2 daalt.

In het geval dat de bediening van klanten van klasse 2 moet herhaald worden – al dan niet met een nieuwe monsterwaarde – impliceert de toegenomen grilligheid van het ondervonden vakantieproces niet alleen dat de wachtlijn van de klanten van klasse 2 gedurende langere periodes aangroeit. Langere uit-periodes impliceren immers ook

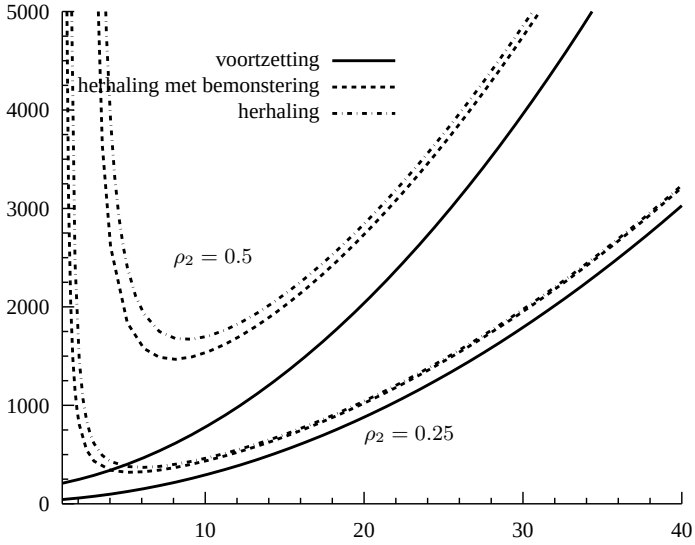


Figuur S.8: Gemiddelde wachttijd van een klant van klasse 2 in functie van de gemiddelde bedieningstijd van de klanten van klasse 1 voor verschillende waarden van de belasting ρ_2 van klasse 2. (2 prioriteitsklassen; klasse 1: Bernoulli-aankomsten, geometrisch verdeelde bedieningstijden, belasting $\rho_1 = 0.2$; klasse 2: Bernoulli-aankomsten, verschoven symmetrisch binomiaal verdeelde bedieningstijden met gemiddelde $\mu_{S_2} = 10$ slots.)

dat er minder bedieningsonderbrekingen zijn. Als de gemiddelde bedieningstijd van de klanten van klasse 1 erg kort is, zullen de lengtes van de door de klasse 2 klanten ondervonden aan- en uit-tijden ook erg kort zijn. Bijgevolg zal de bediening van de klanten van klasse 2 vaak onderbroken worden. Langere bedieningstijden van klanten van klasse 1 – en dus langere aan- en uit-tijden – impliceren dan een verbetering van de performantie van de klanten van klasse 2 omdat hun bediening minder vaak onderbroken wordt. Als de bedieningstijden van de klanten van klasse 1 al erg lang zijn, zal de bediening van klanten van klasse 2 sowieso al weinig onderbroken worden. Een toename van de bedieningstijden van de klanten van klasse 1 – langere aan- en uit-tijden – impliceert dan een afname van de performantie van de klanten van klasse 2. Zoals voor de preëmptieve prioriteitsregeling met voortzetting zullen de klasse 2 klanten steeds langer moeten wachten tijdens steeds langere uit-tijden.

S.3 Andere vakantiemodellen

In tegenstelling tot de vakantiemodellen die in de vorige sectie onderzocht werden, beschouwen we in deze sectie vakantiemodellen waarbij het bedieningsstation rekening houdt met de toestand van het systeem bij de beslissing om al dan niet een vakantie



Figuur S.9: Variantie van de wachttijd van een klant van klasse 2 in functie van de gemiddelde bedieningstijd van de klanten van klasse 1 voor verschillende waarden van de belasting ρ_2 van klasse 2. (2 prioriteitsklassen; klasse 1: Bernoulli-aankomsten, geometrisch verdeelde bedieningstijden, belasting $\rho_1 = 0.2$; klasse 2: Bernoulli-aankomsten, verschoven symmetrisch binomiaal verdeelde bedieningstijden met gemiddelde $\mu_{S_2} = 10$ slots.)

te nemen. I.e., het op vakantie vertrekken kan afhangen van het aantal aanwezige klanten in de buffer, de overblijvende bedieningstijd van de klant die bediend wordt, enzovoort.

Deze sectie vat hoofdstuk 3 van dit proefschrift samen en is grotendeels gebaseerd op de volgende publicaties: [Fiems en Bruneel, 2002a], [Fiems et al., 2002a] en [Fiems en Bruneel, 2003].

S.3.1 Klassieke vakantiemodellen

In de literatuur beschouwt men traditioneel de volgende types vakantiemodellen: Voor vakantiesystemen met exhaustieve bediening (*Engels*: exhaustive vacation system) vertrekt het bedieningsstation enkel op vakantie als er geen enkele klant meer in het systeem aanwezig is. Het bedieningsstation van een wachtlijnsysteem met vakanties en voorzien van een poort (*Engels*: gated vacation system) blijft doorgaan met het bedienen van klanten totdat alle klanten die voor het einde van de laatste vakantie in het systeem aankwamen, bediend zijn en neemt dan een vakantie. In-aantal-limiterende vakantiemodellen (*Engels*: number-limited vacation system) beperken het aantal klanten dat tussen twee opeenvolgende vakanties bediend kan worden. Tijd-limiterende

vakantiemodellen (*Engels*: time-limited vacation system) daarentegen beperken de tijd tussen twee opeenvolgende vakanties. Voor beide limiterende systemen veronderstelt men verder ook dat het bedieningsstation op vakantie vertrekt wanneer er geen klanten meer in het wachtlijnsysteem aanwezig zijn vooraleer de respectievelijke maxima overschreden worden.

Daarnaast onderscheidt men vakantiemodellen met meervoudige en met enkelvoudige vakanties. In een systeem met meervoudige vakanties vertrekt het bedieningsstation onmiddellijk opnieuw op vakantie als er geen klanten in het systeem zijn bij zijn terugkeer van een vakantie. In een systeem met enkelvoudige vakanties daarentegen wacht het bedieningsstation op de eerste klant indien er geen klanten zijn bij zijn terugkeer van vakantie.

Al deze vakantiesystemen werden geanalyseerd door Takagi [1991, 1993]. Zowel discrete-tijd-wachtlijnmodellen als continue-tijd-wachtlijnmodellen met verschillende soorten vakanties werden geanalyseerd. Recente uitbreidingen worden besproken in hoofdstuk 3 van dit proefschrift.

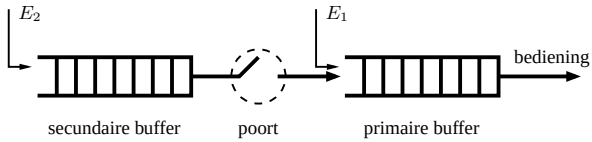
S.3.2 Een vakantiemodel met een poort en exhaustieve bediening

Het eerste model dat we analyseren is een vakantiemodel met een poort en met exhaustieve bediening.

Het model

Het bedieningsstation van het traditionele vakantiemodel voorzien van een poort blijft doorgaan met het bedienen van klanten totdat alle klanten die voor het einde van de laatste vakantie in het systeem aankwamen, bediend zijn. Zo een systeem kan voorgesteld worden als een systeem met twee door een poort gescheiden buffers zoals afgebeeld in figuur S.10. Klanten komen aan in de secundaire buffer en verhuizen gezamenlijk naar de primaire buffer als de poort opent. Deze poort opent op het einde van een vakantie. De primaire buffer werkt als een vakantiesysteem met exhaustieve bediening. I.e., het bedieningsstation vertrekt op vakantie als er geen klanten meer aanwezig zijn in de primaire buffer.

We breiden het klassieke model met een poort nu uit door – naast de aankomsten in de secundaire buffer – ook klanten rechtstreeks in de primaire buffer toe te laten. We noemen dit specifieke vakantiemodel een vakantiemodel met een poort en exhaustieve bediening (*Engels*: gated-exhaustive vacation model). Zo een vakantiesysteem is niet enkel een uitbreiding van het klassieke van een poort voorziene vakantiesysteem, het is ook een uitbreiding van het vakantiesysteem met exhaustieve bediening. Immers, indien we veronderstellen dat er geen aankomsten in de secundaire buffer zijn, zal het



Figuur S.10: Het vakantiemodel met een poort en met exhaustieve bediening.

bedieningsstation enkel vakanties nemen als er geen klanten in het systeem aanwezig zijn. We bekomen in dit geval dus een vakantiesysteem met exhaustieve bediening.

We bepalen in dit proefschrift in het bijzonder de performantie van het vakantiemodel met een poort en met exhaustieve bediening in het geval dat de aankomsten gedurende de opeenvolgende slots in de primaire en de secundaire buffer twee reeksen onderling afhankelijke maar in de tijd onafhankelijke toevalsgrootheden zijn. Het aankomstproces kan dus gekarakteriseerd worden door de gezamenlijke probabiliteitsgenererende functie van het aantal aankomsten in de primaire en de secundaire buffer tijdens een slot. Verder onderstellen we – zoals voorheen – dat de opeenvolgende bedieningstijden van de klanten een reeks onafhankelijke identiek gedistribueerde toevalsgrootheden vormen. De distributie van de bedieningstijden kan willekeurig gekozen worden. Tenslotte vormen de opeenvolgende lengtes van de vakanties ook een reeks onafhankelijke toevalsgrootheden. De distributie van de vakantie kan echter verschillend gekozen worden voor vakanties die onmiddellijk voorafgegaan worden door een andere vakantie en voor vakanties waarvoor dit niet het geval is. Zo een vakantieproces reduceert zich tot een proces met meervoudige vakanties als beide vakantiedistributies gelijk zijn en reduceert zich tot een proces met enkelvoudige vakanties als we veronderstellen dat de lengte van een vakantie die onmiddellijk door een andere vakantie wordt voorafgegaan deterministisch gelijk is aan één slot.

Methode van de toegevoegde variabelen

We analyseren het voorgestelde vakantiesysteem met een poort en exhaustieve bediening met de methode van de toegevoegde variabelen.

In een eerste stap vertalen we de beschrijving van het systeem in een set systeemvergelijkingen. De systeemvergelijkingen relateren de toestand van het systeem op een bepaalde slotgrens aan de toestand van het systeem op de voorafgaande slotgrens. De toestand van het systeem bij het begin van een willekeurig slot kan beschreven worden door de volgende variabelen:

- het aantal klanten in de primaire buffer.
- het aantal klanten in de secundaire buffer.

- het aantal beschikbare bedieningsstations (geen of één).
- het aantal resterende vakantieslots indien het bedieningsstation niet beschikbaar is of de resterende bedieningstijd van de klant die bediend wordt indien het bedieningsstation beschikbaar is.

In een tweede stap gebruiken we de systeemvergelijkingen om de gezamenlijke (partiële) genererende functies van de toestand op een bepaalde slotgrens te relateren aan de gezamenlijke (partiële) genererende functies van de toestand op de voorafgaande slotgrens.

We onderstellen dan dat het systeem stochastisch evenwicht bereikt. Uit het voorafgaande verband tussen de genererende functies van de toestand op opeenvolgende slotgrenzen kunnen we dan de genererende functies van de toestand afleiden in stochastisch evenwicht. We maken hier gebruik van de eigenschappen van genererende functies – met name de normalisatievoorwaarde en het feit dat genererende functies analytisch zijn binnen de eenheidscirkel – om een aantal onbekende functies te bepalen.

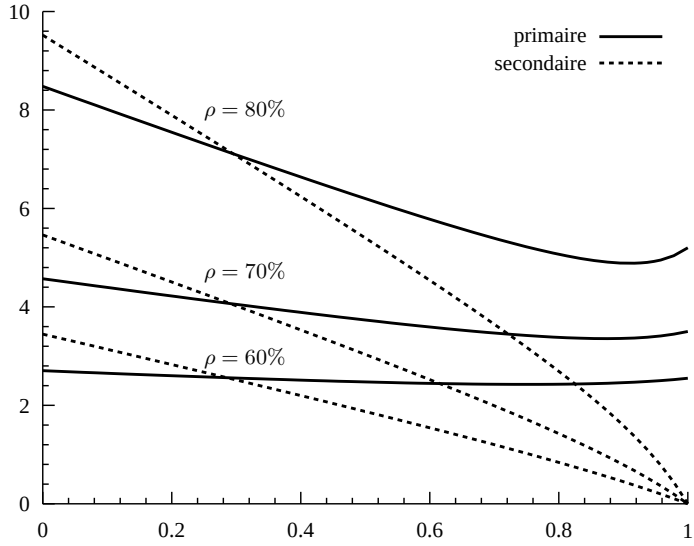
De resultaten laten toe alle mogelijke momenten analytisch te bepalen in functie van één variabele die men numeriek dient te bepalen. We bepalen een set recursieve vergelijkingen die toelaat deze variabele nauwkeurig te bepalen.

Resultaten

De uitdrukkingen voor de (partiële) genererende functies van de toestand van het systeem in stochastisch evenwicht kunnen nu gebruikt worden om uitdrukkingen voor de probabiliteitsgenererende functie van de bufferbezetting op verschillende tijdstippen in stochastisch evenwicht te bepalen alsook voor de probabiliteitsgenererende functies van de wachttijden van klanten die in de primaire en de secundaire buffer aankomen.

In het bijzonder tonen we aan dat de genererende functie van de totale bufferbezetting (het aantal klanten in de primaire en in de secundaire buffer samen) op willekeurige slotgrenzen het product is van de genererende functie van de bufferbezetting op willekeurige slotgrenzen van een wachtlijnsysteem zonder vakanties en van de genererende functie van de bufferbezetting van het hier beschouwde systeem bij het begin van een willekeurig vakantieslot. Deze eigenschap – de stochastische decompositieeigenschap van wachtlijnmodellen met vakanties – geldt voor vrij algemene wachtlijnmodellen met vakanties (zie onder meer Fuhrmann en Cooper [1985]).

Ter illustratie werken we ook enkele numerieke voorbeelden uit. In figuren S.11 en S.12 zetten we bijvoorbeeld de gemiddelde waarde en de variantie van de primaire en secundaire bufferbezetting uit in functie van de fractie x van de aankomsten die in de primaire buffer aankomen bij gegeven systeembelasting ρ. We veronderstellen dat het aantal klanten die in de primaire en in de secundaire buffer aankomen onafhankelijk

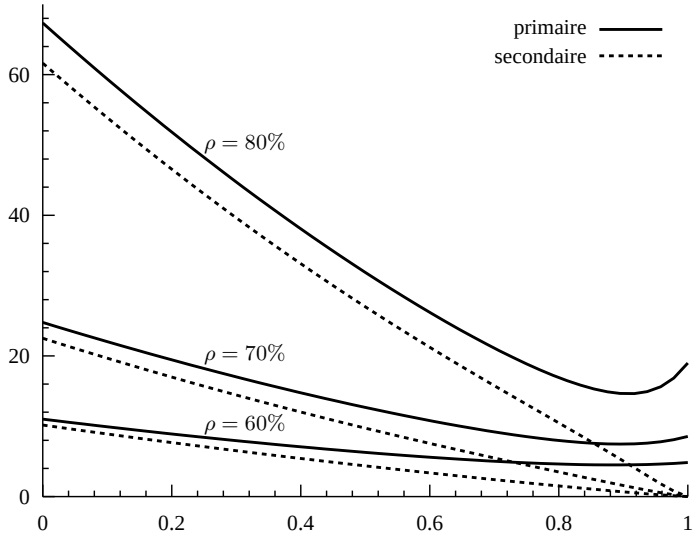


Figuur S.11: Gemiddelde primaire en secundaire bufferbezetting in functie van de fractie x van de aankomsten die in de primaire buffer aankomen voor verschillende waarden van de totale systeembelasting ρ. (Poisson-aankomsten in primaire en secundaire wachtlijn, verschoven geometrisch verdeelde bedieningstijden met gemiddelde $\mu_S = 5$ slots, vakanties van 20 slots.)

van elkaar zijn en dat beiden Poisson-verdeeld zijn. De vakanties duren steeds 20 slots (deterministisch) en de bedieningstijden van de klanten zijn verschoven geometrisch verdeeld met gemiddelde waarde $\mu_S = 5$ slots. Voor $x = 0$ komen er enkel klanten in de secundaire buffer aan. We krijgen het klassieke vakantiemodel voorzien van een poort. Voor $x = 1$ komen alle klanten in de primaire buffer aan. We krijgen dus een vakantiemodel met exhaustieve bediening. Voor toenemende waarden van x komen er steeds minder klanten in de secundaire buffer aan. De gemiddelde secundaire bufferbezetting en de bijhorende variantie dalen daarom voor toenemende waarden van x. Minder verwacht is dat ook de primaire bufferbezetting alsook de bijhorende variantie dalen (althans voor voldoende kleine x) voor toenemende waarden van x. Een gedeeltelijke verklaring ligt in het feit dat het aankomstproces in de primaire buffer minder grillig is naarmate er meer klanten rechtstreeks in deze buffer aankomen. Typisch voor dit wachtlijnmodel echter is dat de bekomen numerieke resultaten moeilijk sluitend intuïtief te verklaren zijn.

S.3.3 Een raamwerk voor vakantiemodellen zonder poorten

Naast het exhaustieve systeem met een poort beschouwen we in deze sectie ook een raamwerk voor de prestatieanalyse van discrete-tijd-wachtlijnmodellen met vakanties en zonder poorten.



Figuur S.12: Variantie van de primaire en secundaire bufferbezetting in functie van de fractie x van de aankomsten die in de primaire buffer aankomen voor verschillende waarden van de totale systeembelasting ρ. (Poisson-aankomsten in primaire en secundaire wachtlijn, verschoven geometrisch verdeelde bedieningstijden met gemiddelde $\mu_S = 5$ slots, vakanties van 20 slots.)

Het model

Zoals steeds beschouwen we een discrete-tijd-wachtlijnmodel met een oneindige buffercapaciteit. We maken de gebruikelijke onderstellingen betreffende het aankomstproces en de bedieningstijden. I.e., de aantallen aankomsten van klanten gedurende de opeenvolgende slots vormen een reeks onafhankelijke identiek gedistribueerde niet-negatieve toevalsgrootheden en de bedieningstijden van deze klanten vormen een reeks onafhankelijke identiek gedistribueerde positieve toevalsgrootheden.

Het vakantieproces bevindt zich in 1 van de N mogelijke (Markoviaanse) staten tijdens slots waar het bedieningsstation beschikbaar is. Op het einde van zo een slot neemt het bedieningsstation een vakantie. Na de vakantie is het bedieningsstation opnieuw beschikbaar gedurende een slot en bevindt het zich in 1 van de N mogelijke staten. Op het einde van dit slot neemt het station terug een vakantie, enz. We laten "vakanties" van nul slots toe. In dat geval blijft het bedieningsstation beschikbaar. De kans dat het bedieningsstation na een bepaald slot een vakantie van een zekere duur neemt en daarna naar een bepaalde staat overgaat hangt af van de staat waarin het vakantieproces zich gedurende dit slot bevindt alsook van de "toestand" waarin het systeem zich bevindt. We laten toe verschillende probabiliteiten te definiëren voor de volgende toestanden:

- er is een klant in bediening en deze verlaat het systeem op het einde van het slot.

Er zijn nog andere klanten in het systeem na het vertrek van deze klant.

- er is een klant in bediening en deze verlaat het systeem op het einde van het slot. Er zijn geen andere klanten meer in het systeem na het vertrek van deze klant.

- er is een klant in bediening en deze verlaat het systeem niet op het einde van het slot.

- er zijn geen klanten aanwezig in het systeem.

Aangezien we toelaten dat het bedieningsstation op vakantie vertrekt gedurende het bedienen van een klant, kunnen we opnieuw verschillende werkingsmodi bestuderen. Zoals in sectie S.2, onderzoeken we de performantie van de CAI modus, de RAI modus en de RAI,wr modus.

De analyse

We gaan hier min of meer op dezelfde wijze te werk als in sectie S.2. We bepalen eerst de partiële conditionele genererende functies van de afwerkingstijd van een klant. De afwerkingstijd begint bij het begin van het slot waar de klant voor het eerst bediend wordt en eindigt op het einde van het slot waar de klant het systeem verlaat. De partiële conditionele genererende functie van de afwerkingstijd heeft de volgende vorm:

$$C^{(ij)}(z) = \mathrm{E}\left[z^C | Q_{\text{laatste}} = j, Q_{\text{eerste}} = i\right] \Pr\left[Q_{\text{laatste}} = j\right], \tag{S.8}$$

met $i, j \in \{1 \ldots N\}$. In de bovenstaande formule duidt C de afwerkingstijd van de klant aan en duiden Q_{laatste} en Q_{eerste} de toestand van het vakantieproces aan gedurende respectievelijk het eerste en het laatste slot van de afwerkingstijd.

Bij de analyse combineren we het gebruik van genererende functies met het gebruik van matrices, i.e., we manipuleren matrices met partiële conditionele genererende functies als elementen. Het gebruik van matrices vereenvoudigt de notatie aanzienlijk. De eindigheid van de toestandsruimte van het vakantieproces is hier niet vreemd aan.

Eenmaal we de karakteristieken van de afwerkingstijden bepaald hebben, bepalen we de probabiliteitsgenererende functies van de bufferbezetting op vertrektijdstippen en op willekeurige slotgrenzen. Het gebruik van de afwerkingstijden laat opnieuw toe dat de eigenlijke wachtlijnanalyse simultaan voor de verschillende werkingsmodi kan uitgevoerd worden.

Bijzondere gevallen

Het beschouwde wachtlijnmodel laat toe de performantie te bepalen van verschillende klassieke vakantiemodellen, zowel systemen zonder bedieningsonderbrekingen als systemen met bedieningsonderbrekingen.

Systemen zonder bedieningsonderbrekingen In het geval dat er geen bedieningsonderbrekingen zijn, dienen we slechts de afwerkingstijden voor één enkele werkingsmodus te bepalen. Er zijn immers geen onderbrekingen. Al de beschouwde werkingsmodi (CAI, RAI en RAI,wr) leiden tot dezelfde resultaten. In het bijzonder bepalen we de probabiliteitsgenererende functie van de bufferbezetting op willekeurige slotgrenzen voor de volgende vakantiesystemen:

- het vakantiesysteem met exhaustieve bediening en meervoudige vakanties.
- het vakantiesysteem met exhaustieve bediening en enkelvoudige vakanties.
- het in-aantal-gelimiteerde vakantiesysteem met meervoudige vakanties.
- het niet-preëmptieve tijdsgelimiteerde vakantiesysteem met meervoudige vakanties.

Systemen met bedieningsonderbrekingen Indien we wel bedieningsonderbrekingen toelaten, leiden de verschillende werkingsmodi over het algemeen tot verschillende resultaten. We bepalen in het bijzonder de probabiliteitsgenererende functie van de bufferbezetting op willekeurige slotgrenzen voor de volgende vakantiesystemen:

- het systeem met willekeurige vakanties zoals dit reeds in sectie S.2 werd behandeld.
- het preëmptieve tijdsgelimiteerde vakantiesysteem (preëmptief met voortzetting, preëmptief met herhaling, preëmptief met herhaling en opnieuw bemonsteren) met verschoven geometrisch verdeelde tijdslimiet.

Naast deze “traditionele” vakantiesystemen laat het onderzochte model natuurlijk ook toe de prestatie van meer gecompliceerde vakantiemodellen te onderzoeken. Bijvoorbeeld modellen met gecorreleerde vakanties, onderbrekingsmodellen met meer algemene aan-tijden, enzovoort.

Bibliography

A. Alfa. A discrete MAP/PH/1 queue with vacations and exhaustive time-limited service. *Operations Research Letters*, 18(1):31–40, 1995.

M. Ali, X. Zhang, and J. Hayes. A discrete-time queueing analysis of the wireless ATM multiplexing system. In P. Lorenz, editor, *Networking – ICN 2001 (Part I)*, volume 2093 of *Lecture notes in computer science*, pages 429–438, Colmar, France, July 2001. Springer.

J. Artalejo. Analysis of an M/G/1 queue with constant repeated attempts and server vacations. *Computers and Operations Research*, 24(6):493–504, 1997.

B. Avi-Itzhak and P. Naor. Some queuing problems with the service station subject to breakdown. *Operations Research*, 11(3):303–319, 1963.

T. Birdsall, M. Ristenbatt, and S. Weinstein. Analysis of asynchronous time multiplexing of speech sources. *IRE Transactions on Communication Systems*, CS-10: 390–397, 1962.

O. Boxma and U. Yechiali. An M/G/1 queue with multiple types of feedback and gated vacations. *Journal of Applied Probability*, 34(3):773–784, 1997.

H. Bruneel. Analysis of buffer behaviour for an integrated voice-data system. *Electronics Letters*, 19(2):72–74, 1983a.

H. Bruneel. Buffers with stochastic output interruptions. *Electronics Letters*, 19(18): 735–737, 1983b.

H. Bruneel. A general model for the behaviour of infinite buffers with periodic service opportunities. *European Journal of Operational Research*, 16:98–106, 1984a.

H. Bruneel. On buffers with stochastic input and output interruptions. *International Journal of Electronics and Communications (AEÜ)*, 38(4):265 –271, 1984b.

H. Bruneel. A discrete-time queueing system with a stochastic number of servers subjected to random interruptions. *Opsearch*, 22(4):215–231, 1985.

H. Bruneel. A general treatment of discrete-time buffers with one randomly interrupted output line. *European Journal of Operational Research*, 27(1):67–81, 1986.

H. Bruneel. Performance of discrete-time queuing systems. *Computers and Operations Research*, 20:303–320, 1993.

H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic Publishers, 1993.

H. Bruneel, J. De Vriendt, and C. Ysebaert. Receiver buffer behavior for the selective-repeat ARQ protocol. *Computer Networks and ISDN Systems*, 19:129–142, 1990.

J. Chiarawongse, M. Srinivasan, and T. Teorey. The M/G/1 queuing system with vacations and timer-controlled service. *IEEE Transactions on Communications*, 42:1846–1855, 1994.

B. Choi, B. Kim, and S. Choi. An M/G/1 queue with multiple types of feedback, gated vacations and FCFS policy. *Computers and Operations Research*, 30(9): 1289–1309, 2003.

J. Cohen. *The single server queue*. North-Holland Pub. Co., Amsterdam, 1969.

R. Cooper. *Introduction to queueing theory*. Elsevier North-Holland, Inc., New York, 1981.

R. Cooper, S. Niu, and M. Srinivasan. Some reflections on the renewal-theory paradox in queueing theory. *Journal of Applied Mathematics and Stochastic Analysis*, 11(3): 355–368, 1997.

R. Cowan. An extension of Tanner's results on uncontrolled intersections. *Queueing Systems*, 1:249–263, 1987.

D. Cox. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society*, 51: 433–441, 1955.

H. Daduna. *Queueing networks with discrete time scale: Explicit expressions for the steady state behavior of discrete time stochastic networks*. Springer Verlag, 2001.

B. Doshi. Queueing systems with vacations – a survey. *Queueing Systems*, 1:29–66, 1986.

I. Eliazar and U. Yechiali. Polling under the randomly timed gated regime. *Stochastic Models*, 14:79–93, 1998a.

I. Eliazar and U. Yechiali. Randomly timed gated queueing systems. *SIAM Journal on Applied Mathematics*, 59:423–441, 1998b.

A. Federgruen and L. Green. Queueing systems with service interruptions. *Operations Research*, 34(5):752–768, 1986.

D. Fiems and H. Bruneel. Discrete-time queueing systems with vacations governed by geometrically distributed timers. In *Proceedings of Africom 2001, 5th International Conference on Communication Systems*, Cape Town, South Africa, May 2001.

D. Fiems and H. Bruneel. Analysis of a discrete-time queueing system with timed vacations. *Queueing Systems*, 42(3):243–254, 2002a.

D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002b.

D. Fiems and H. Bruneel. Discrete-time queues with correlated vacations. In *International Teletraffic Congress (ITC-18)*, pages 581–590, Berlin, Germany, Sept. 2003.

D. Fiems, S. De Vuyst, and H. Bruneel. The combined gated-exhaustive vacation system in discrete time. *Performance Evaluation*, 49(1–4):225–237, 2002a.

D. Fiems, B. Steyaert, and H. Bruneel. Performance evaluation of CAI and RAI transmission modes in a GI-G-1 queue. *Computers and Operations Research*, 28 (13):1299–1313, 2001.

D. Fiems, B. Steyaert, and H. Bruneel. Analysis of a discrete-time GI-G-1 queueing model subjected to bursty interruptions. *Computers and Operations Research*, 30 (1):139–153, 2002b.

D. Fiems, B. Steyaert, and H. Bruneel. Randomly interrupted GI-G-1 queues, service strategies and stability issues. *Annals of Operations Research*, 112:171–183, 2002c.

D. Fiems, B. Steyaert, and H. Bruneel. Discrete-time queues with generally distributed service times and renewal-type server interruptions. *Performance Evaluation*, 55 (3–4):277–298, 2003.

D. Fiems, J. Walraevens, and H. Bruneel. The discrete-time gated vacation queue revisited. *International Journal of Electronics and Communications (AEÜ)*, (to appear), 2004.

A. Frey and Y. Takahashi. A note on an M/GI/1/N queue with vacation time and exhaustive service discipline. *Operations Research Letters*, 21(2):95–100, 1997.

S. Fuhrmann and R. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33(5):1117 –1129, 1985.

D. Gaver Jr. A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society*, B24:73–90, 1962.

N. Georganas. Buffer behavior with Poisson arrivals and bulk geometric output processes. *IEEE Transactions on Communications*, 24(8):938–940, 1976.

T. Heines. Buffer behavior in computer communication systems. *IEEE Transactions on Communications*, 28:573–576, 1979.

J. Hsu. Buffer behavior with Poisson arrival and geometric output processes. *IEEE Transactions on Communications*, 22:1940–1941, 1974.

J. Hunter. *Mathematical Techniques of Applied Probability, Volume 1*. Operations Research and Industrial Engineering. Academic Press, New York, 1983a.

J. Hunter. *Mathematical Techniques of Applied Probability, Volume 2*. Operations Research and Industrial Engineering. Academic Press, New York, 1983b.

O Ibe and K. Trivedi. Two queues with alternating service and server breakdown. *Queueing Systems*, 7(3-4):253–268, 1990.

V. Inghelbrecht, K. Laevens, H. Bruneel, and B. Steyaert. Queueing of fixed-length messages in the presence of server interruptions. In *Proceedings Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2k*, Vancouver, Canada, July 2000.

F. Ishizaki, T. Takine, and T. Hasegawa. Analysis of a discrete-time queue with a gate. In J. Labetoulle and J. Roberts, editors, *Proceedings of ITC 14*, pages 169 –178. Elsevier Science B.V., 1994.

F. Ishizaki, T. Takine, and T. Hasegawa. Analysis of a discrete-time queue with gated priority. *Performance Evaluation*, 23(2):121–143, 1995.

J. Kim and M. Krunz. Delay analysis of selective repeat ARQ for a Markovian source over a wireless channel. *IEEE Transactions on Vehicular Technology*, 49(5):1968–1981, 2000.

L. Kleinrock. *Queueing systems, Volume I: theory*. John Wiley & Sons, New York, 1975.

L. Kleinrock. *Queueing systems, Volume II: computer applications*. John Wiley & Sons, New York, 1976.

K. Laevens and H. Bruneel. Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. *European Journal of Operational Research*, 85:161–177, 1995.

K. Laevens and H. Bruneel. Discrete-time multiserver queues with priorities. *Performance Evaluation*, 33(4):249–275, 1998.

R. LaMaire. An M/G/1 vacation model of an FDDI station. *IEEE Journal on Selected Areas in Communications*, 9(2):257–264, 1991.

A. Law and W. Kelton. *Simulation, modeling & analysis*. Industrial Engineering Series. McGraw-Hill international editions, second edition, 1991.

D. Lee. Analysis of a single server queue with semi-Markovian service interruption. *Queueing Systems*, 27(1–2):153–178, 1997a.

T. Lee. A simple approach for analyzing feedback vacation queues with levy input process. *European Journal of Operational Research*, 96(2):299–316, 1997b.

Y. Lee. Discrete-time $Geo^X/G/1$ queue with preemptive resume priority. *Mathematical and Computer Modelling*, 34(3–4):243–250, 2001.

Y. Lee and K. Lee. Discrete-time $Geo^X/G/1$ queue with preemptive repeat different priority. *Queueing Systems*, 44(4):399–411, 2003.

K. Leung and M. Eisenberg. A single-server queue with vacations and gated time-limited service. *IEEE Transactions on Communications*, 38(9):1454 –1462, 1990.

K. Leung and M. Eisenberg. A single-server queue with vacations and non-gated time-limited service. *Performance Evaluation*, 12:115–125, 1991.

K. Leung and D. Lucantoni. Two vacation models for token-ring networks where service is controlled by timers. *Performance Evaluation*, 20:165–184, 1994.

H. Li and Y. Zhu. Analysis of M/G/1 queues with delayed vacations and exhaustive service discipline. *European Journal of Operational Research*, 92(1):125–134, 1996.

J. Little. A proof for the queueing formula: $L = \lambda W$. *Operations Research*, 9: 383–387, 1961.

T. Meisling. Discrete-time queueing theory. *Operations Research*, 6(1):96–105, 1958.

P. Nain. Queueing systems with service interruptions: an approximate model. *Performance Evaluation*, 3(2):123–129, 1983.

M. Neuts. Matrix-analytic methods in queueing theory. *European Journal of Operational Research*, 15:2–12, 1983.

R. Núñez Queija. Sojourn times in a processor sharing queue with service interruptions. *Queueing Systems*, 34(1–4):351–386, 2000.

D. Perry and M. Posner. A correlated M/G/1-type queue with randomized server repair and maintenance modes. *Operations Research Letters*, 26(3):137–147, 2000.

B. Powell and B. Avi-Itzhak. Queuing systems with enforced idle times. *Operations Research*, 15(6):1145–1156, 1967.

I. Rubin and Z. Tsai. Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems. *IEEE Transactions on Information Theory*, 35 (2):637–647, 1989.

L. Servi and S. Finn. M/M/1 queues with working vacations (M/M/1/WV). *Performance Evaluation*, 50(1):41–52, 2002.

N. Shacham and D. Towsley. Resequencing delay and buffer occupancy in selective repeat ARQ with multiple receivers. *IEEE Transactions on Communications*, 39 (6):928–937, 1991.

M. Shomrony and U. Yechiali. Burst arrival queues with server vacations and random timers. *Mathematical Methods of Operations Research*, 53(1):117–146, 2001.

S. Stidham. Analysis, design, and control of queueing systems. *Operations Research*, 50:197–216, 2002.

L. Takács. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.

H. Takagi. *Queueing Analysis; A foundation of performance evaluation, volume 1: Vacation and priority systems, part 1*. Elsevier Science Publishers, 1991.

H. Takagi. *Queueing Analysis; A foundation of performance evaluation, volume 3: Discrete-time systems*. Elsevier Science Publishers, Amsterdam, 1993.

H. Takagi. M/G/1//N queues with server vacations and exhaustive service. *Operations Research*, 42(5):926–939, 1994.

M. Takahashi, H. Osawa, and T. Fujisawa. $Geo^{[X]}/G/1$ retrial queue with non-preemptive priority. *Asia-Pacific Journal of Operational Research*, 16(2):215–234, 1999.

T. Takine and B. Sengupta. A single server queue with service interruptions. *Queueing Systems*, 26:285–300, 1997.

T. Takine, B. Sengupta, and T. Hasegawa. An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. *IEEE Transactions on Communications*, 42(2/3/4), 1994.

K. Thiruvengadam. Queuing with breakdowns. *Operations Research*, 11(1):62–71, 1963.

D. Towsley. A statistical analysis of ARQ protocols operating in a nonindependent error environment. *IEEE Transactions on Communications*, 29(7):971–981, 1981.

D. Towsley and J. Wolf. On the statistical analysis of queue lengths and waiting times for statistical multiplexers with ARQ retransmission schemes. *IEEE Transactions on Communications*, 27(4):693–702, 1979.

F. Van der Duyn Schouten and S. Vanneste. Maintenance optimization of a production system with buffer capacity. *European Journal of Operational Research*, 82:232–338, 1995.

N. Van Dijk. Simple bounds for queueing systems with breakdowns. *Performance Evaluation*, 8(2):117–128, 1988.

J. Walraevens, B. Steyaert, and H. Bruneel. Delay characteristics in discrete-time GI-G-1 queues with non-preemptive priority queueing discipline. *Performance Evaluation*, 50(1):53–75, 2002.

J. Walraevens, B. Steyaert, and H. Bruneel. Analysis of a preemptive repeat priority buffer with resampling. In *Proceedings of the International Network Optimization Conference*, Evry, France, October 2003a.

J. Walraevens, B. Steyaert, and H. Bruneel. Delay analysis of a discrete-time non-preemptive priority buffer with 3 traffic classes. In *Proceedings of the 7th WSEAS International Multiconference CSCC*, Corfu, July 2003b.

J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline. *European Journal of Operational Research*, (to appear), 2003c.

J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a single-server ATM queue with a priority scheduling. *Computers and Operations Research*, 30 (12):1807–1829, 2003d.

P. Wang, J. Gray, and M. Scott. Quality-related measures of unreliable machines with preemptive maintenance. *Computers and Operations Research*, 23(10):981–996, 1995.

H. White and L. Christie. Queuing with preemptive priorities or with breakdown. *Operations Research*, 6(1):79–95, 1958.

W. Whitt. A review of $L = \lambda W$ and extensions. *Queueing Systems*, 9:235–268, 1991.

C. Woodside and E. Ho. Engineering calculation of overflow probabilities in buffers with Markov-interrupted service. *IEEE Transactions on Communications*, 35(12): 1272–1277, 1987.

M. Woodward. *Communication and Computer Networks: Modelling with discrete-time queues*. Wiley-IEEE Computer Society Press, 1993.

O. Yang and J. Mark. Performance analysis of integrated services on a single server system. *Performance Evaluation*, 11:79–92, 1990.

M. Yoshimoto, T. Takine, Y. Takahashi, and T. Hasegawa. Waiting time and queue length distributions for go-back-N and selective-repeat ARQ protocols. *IEEE Transactions on Communications*, 41(11):1687–1693, 1993.