Analyse van generische discrete-tijd-buffermodellen
met grillige aankomstpatronen

Analysis of Generic Discrete-Time Buffer Models
with Irregular Packet Arrival Patterns

Bart Steyaert

Promotor: prof. dr. ir. H. Bruneel
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen
Academiejaar 2007 - 2008

UNIVERSITEIT
GENT

ISBN 978-90-8578-187-5
NUR 919, 992
Wettelijk depot: D/2008/10.500/6

Dankwoord

Ik zou graag iedereen willen bedanken die in de loop der jaren, van dichtbij of veraf, betrokken was of is bij mijn onderzoek en het schrijven van dit proefschrift.

Daarbij denk ik uiteraard eerst en vooral aan mijn promotor, Prof. H. Bruneel, die mij met veel enthousiasme zowat twintig jaar geleden heeft laten kennismaken met het mij toen onbekende onderzoeksdomein van de wachtlijntheorie, die mij daarna heeft gestimuleerd om zelf mijn eerste stapjes te zetten in dit domein, en zonder wiens voortdurende steun dit werk waarschijnlijk (nog) niet tot stand zou zijn gekomen. Mijn appreciatie gaat dan ook een stuk verder dan het zuiver wetenschappelijke.

Van in het begin was er bovendien de (financiële) ondersteuning van Alcatel-Bell (nu Alcatel-Lucent, Antwerpen) – later samen met het IWT – die dit onderzoek heeft mogelijk gemaakt. In het bijzonder denk ik aan dr. G.H. Petit, met wie het altijd plezierig samenwerken was, en die met zijn 'vraagjes' ondermeer de aanzet vormde voor mijn onderzoek van de eerste wachtlijnmodellen.

Wetenschappelijk onderzoek is altijd opnieuw een leerproces, soms met vallen en opstaan; vandaar dat de interactie met de collega's van de SMACS-onderzoeksgroep onontbeerlijk was, en nog altijd is. Het zijn er ondertussen teveel om op te noemen, maar ik denk in ieder geval met veel plezier aan de samenwerking met Yijun, Koenraad, Sabine, Véronique, Danny, Joris, Dieter, en alle anderen.

Vanzelfsprekend zou ik ook mijn vader en moeder, broers en zus, willen bedanken voor hun steun gedurende al die jaren, en de vrienden, wier aanhoudende vraag 'hoe-zit-dat-nu-met-dat-doctoraat?' mij uiteindelijk zover heeft gekregen.

En dan is er natuurlijk ook nog de liefdevolle steun van Maayken, die zo onmisbaar geworden is.

Bart Steyaert, oktober 2007

‘How did he send the message? I’ve been right with him almost all the time.’
‘He sent it by the usual means,’ Iff shrugged. ‘A P2C2E.’
‘And what is that?’
‘Obvious,’ said the Water Genie with a wicked grin.
‘It’s a Process Too Complicated To Explain.’
S. Rushdie, Haroun and the Sea of Stories

Samenvatting

Deze scriptie geeft een overzicht van een deel van het onderzoekswerk dat ik, sinds mijn beginjaren bij de SMACS-onderzoeksgroep, heb verricht met betrekking tot het modelleren en analyseren van wachtlijnsystemen. Een wachtlijnsysteem kan in het algemeen conceptueel worden omschreven als elke logische of fysische entiteit waar klanten binnenkomen, eventueel een tijd moeten wachten alvorens bediend te worden, waarna ze het systeem verlaten. Wachtlijnsystemen komen we dagdagelijks tegen in onze omgeving; denken we maar aan de wachtrij bij de bakker of de kassa van het grootwarenhuis, met de auto staan wachten in de file, ··· .

Het (analytisch) onderzoek van dergelijke systemen gebeurt veelal aan de hand van een passend mathematisch/stochastisch model, dat de manier beschrijft waarop klanten het systeem binnentreden (het *aankomstproces*), de grootte van de bedieningstijden en het aantal bedieningsstations, het aantal beschikbare plaatsen in de wachtrij (de *buffergrootte*), en de volgorde waarin klanten worden bediend. De wachtlijnmodellen die hier aan bod komen werden ontwikkeld en bestudeerd in het kader van de performantiestudie van buffers die voorkomen in de componenten van telecommunicatienetwerken, zoals ATM schakelelementen, buffers in pakketgebaseerde (zoals IP) netwerken, etc. ··· . In een dergelijke omgeving zijn de 'klanten' de digitale datapakketten die over het netwerk worden verstuurd, en de buffers die in de knopen van een dergelijk netwerk geïmplementeerd worden, dienen voor de (tijdelijke) opvang van de datapakketten die, om welke reden dan ook – maar die in vele gevallen te maken heeft met contentie met andere datapakketten die wedijveren voor de beschikbare bandbreedte – niet onmiddellijk kunnen doorgestuurd worden naar de eerstvolgende bestemming. Door het digitale, en dus discrete, karakter van deze pakketten, worden dergelijke buffersystemen veelal dan ook beschreven aan de hand van een *discrete-tijd* model. Hierbij is de tijd geen continue grootheid, maar wordt de tijdsas ingedeeld in intervallen van gelijke grootte, *slots* genaamd, en wordt het discrete karakter van de tijdsparameter dus gekarakteriseerd door de volgnummers van de opeenvolgende slots. De bedieningstijd, of grootte, van een pakket is dan een maat voor zijn *transmissietijd* (uitgedrukt in aantal slots), en de 'bedieningsstations' zijn de uitgangslijnen of –kanalen van de buffer in de netwerkknoop.

De grootheden die van belang zijn bij het bestuderen van een dergelijk systeem hebben grosso modo betrekking op het aantal pakketten die de buffer bevat (de *bufferbezetting*), de tijd die een pakket in het buffersysteem doorbrengt (de *verblijftijd*) en het *verliesproces* der pakketten die worden geweigerd omdat de buffer reeds volledig gevuld is op het ogenblik dat een nieuw pakket zich aanbiedt. Aangezien zowel het aankomstproces als de pakketgrootte

worden beschreven door middel van een discreet stochastisch proces, wat betekent dat ze gekarakteriseerd worden door één of meerdere *toevalsgrootheden* (en de daarmee geassocieerde *probabiliteitsdistributie*(s)), zijn de hierboven vermelde grootheden eveneens discrete toevalsgrootheden, met als gevolg dat we vooral geïnteresseerd zijn in performantiematen zoals hun momenten – en in het bijzonder de gemiddelde waarde en variantie – (het asymptotisch gedrag van) hun massafunctie, $\cdots$. Wat de pakket-verblijftijden betreft, beperken we ons (grotendeels) tot de situatie waarbij pakketten de wachtrij vervoegen, en dus ook bediend worden, in hun volgorde van aankomst.

Het uiteindelijke doel van dit soort studies is om de nodige formules op te stellen die ons toelaten om de invloed van de systeemparameters op deze grootheden in het gemodelleerde systeem, te bestuderen en in te schatten, en om op basis van deze resultaten dimensioneringsregels op te stellen die (hopelijk) nuttig zijn bij de daadwerkelijke implementatie ervan. De oplossingsmethode die onze voorkeur geniet, maakt intensief gebruik van *probabiliteitsgenererende functies* (pgfs) om de relevante toevalsgrootheden en hun onderlinge verbanden te beschrijven. In de loop van deze scriptie zal duidelijk worden dat deze aanpak een vrij grondige kennis vereist van een aantal mathematische en stochastische analysetechnieken, maar heeft als belangrijk voordeel dat alle resultaten (in de mate van het mogelijke) rechtstreeks worden uitgedrukt als functie van de systeemparameters, wat dus nauw aansluit bij de hierboven geformuleerde doelstelling.

Na een inleidende bespreking in Hoofdstuk I van een aantal begrippen en onderstellingen die van belang zijn bij de studie van deze modellen en waarvan er een aantal reeds summier werden vermeld, bekijken we in Hoofdstuk II een relatief eenvoudig wachtlijnmodel, waarbij aankomsten van pakketten worden gegenereerd door N (identieke) bronnen, en het aankomstproces wordt beschreven met behulp van een rij van toevalsgrootheden die *statistisch onafhankelijk* en *identisch gedistribueerd* worden ondersteld, wat betekent dat één enkele toevalsgrootheid volstaat om het volledige aankomstproces der pakketten te beschrijven. De pakketgrootte wordt gelijk aan 1 slot ondersteld, terwijl het bedieningsstation over c (≥ 1) uitgangskanalen beschikt. In eerste instantie wordt een oneindige buffergrootte beschouwd, wat een aantal voordelen en vereenvoudigingen met betrekking tot de wiskundige analyse met zich meebrengt. Voor dit model beschouwen we de analyse van de bufferbezetting en de verblijftijden van de pakketten met behulp van de hierboven aangehaalde analysetechniek, wat leidt tot uitdrukkingen voor de gemiddelde waarde en variantie van de grootheden, en het asymptotisch gedrag van hun massafunctie (ook wel ‘staartdistributie’ genoemd) op basis van de ‘dominante pool’ benaderingsmethode. In dit deel van het hoofdstuk wordt in feite de basis gelegd voor een aantal berekeningsmethodes en oplossingstechnieken die ook in het vervolg van de scriptie hun deugdelijkheid zullen bewijzen, mits de nodige uitbreidingen en aanpassingen. Daarna richten we onze aandacht op

een buffersysteem met eindige opslagcapaciteit, waarvoor we het verliesproces in detail bestuderen. Meer specifiek slagen we erin om de relevante grootheden die betrekking hebben op een buffer met eindige opslagcapaciteit uit te drukken als functie van de performantiematen voor een buffer met een oneindige opslagruimte, wat leidt tot een efficiënte methode om deze te berekenen.

De onderstelling dat de aankomsten van pakketten gedurende opeenvolgende slots op een onafhankelijke manier worden gegenereerd is niet in alle omstandigheden even realistisch. Vandaar dat we in Hoofdstuk III een algemeen kader schetsen, waarbij afhankelijkheid – of correlatie – kan worden ingebouwd in het model dat het aantal aankomsten beschrijft van pakketten die gedurende opeenvolgende slots door een bron gegenereerd worden. Dit gebeurt met behulp van een zogenaamd *discrete-batch Markovian arrival process* (D-BMAP), waarbij een onderliggende Markov-keten met L toestanden het aantal aankomsten van pakketten in een slot stuurt, daarbij rekening houdend met de toestand van de Markov-keten in het huidige en het voorgaande slot. Een dergelijk aankomstproces heeft een heel grote vrijheidsgraad om afhankelijkheid in het aankomstproces te beschrijven, en is bij herhaling aangewend om 'realistische' aankomstpatronen, zoals de pakketstromen die voortkomen uit digitale videobronnen of telefonie, te karakteriseren. In geval van N identieke (*homogene*) bronnen leiden we opnieuw uitdrukkingen af voor de gemiddelde waarde en variantie van de bufferbezetting. Bovendien beschrijven we ook een efficiënt algoritme om de staartdistributie te berekenen, dat meerdere polen van de pgf van de bufferbezetting in rekening brengt. Ten slotte stellen we ook een verband op tussen de (distributie van) enerzijds de bufferbezetting en anderzijds de verblijftijden der pakketten, dat zich niet beperkt tot een D-BMAP en dus algemeen toepasbaar is. Dit alles wordt rijkelijk geïllustreerd aan de hand van numerieke voorbeelden, wat ons toelaat om enig inzicht te verwerven in de werking en onderliggende mechanismen van een dergelijk buffersysteem. Zo zal bijvoorbeeld blijken dat de 'burst factor', die een karakteristiek is van het aankomstproces, in grote mate de performantie van het systeem bepaalt, in het bijzonder met betrekking tot de momenten van de bufferbezetting en de verblijftijden. Deze resultaten worden vervolgens verder uitgebreid naar het geval van niet-identieke (*heterogene*) bronnen. Daarbij beschouwen we twee mogelijke manieren waarop pakketten die tijdens eenzelfde slot de wachtrij vervoegen kunnen worden geordend : enerzijds kunnen alle pakketten als ononderscheidbaar worden beschouwd en dus in arbitraire volgorde in de wachtrij gestopt worden (*arbitrary ordering*, AO), en anderzijds kan de ordening rekening houden met de klasse waartoe een pakket behoort (*fixed-order-by-class*, FOC). Ook in dit geval slagen we erin om uitdrukkingen op te stellen die ons toelaten om de belangrijkste performantiematen met betrekking tot de bufferbezetting, en de verblijftijd van een pakket van type k, te berekenen, al dient gezegd dat dit in het geval van AO behoorlijk wat extra numerieke berekening vergt in vergelijking met FOC. Dit hoofdstuk wordt tenslotte beëindigd met een *case study*, waarbij we een netwerkknoop beschouwen die gepakketiseerd

telefoonverkeer ondersteunt, waarbij verschillende scenario's in detail worden geëvalueerd. De bedoeling is om aan de hand van dit voorbeeld de efficiëntie en bruikbaarheid te illustreren van de formules en resultaten die in dit hoofdstuk tot stand kwamen.

In Hoofdstuk IV tenslotte beschouwen we de uitbreiding van het model uit het voorgaande hoofdstuk naar de situatie waarbij de pakket-transmissietijden niet langer constant en gelijk aan 1 slot zijn, maar een algemene verdeling hebben. We beperken ons daarbij tot het geval van 1 uitgangskanaal, aangezien een analytische aanpak van dergelijke modellen met meerdere uitgangskanalen tot hiertoe nog niet tot een bevredigende oplossing heeft geleid. De analyse van de bufferbezetting, opnieuw gebaseerd op een intensief en adequaat aanwenden van probabiliteitsgenererende functies en hun eigenschappen, maakt gebruik van de *supplementaire variabele* methode. We bekomen uiteindelijk opnieuw uitdrukkingen voor de gemiddelde waarde en variantie, en staartdistributie, van de bufferbezetting en de pakket verblijftijden, zowel in het geval van een homogeen, als een heterogeen aankomstproces.

Tot slot zouden we willen benadrukken dat, ondanks de eerder 'wiskundige' aard van dit werk, het niet de bedoeling is om voor elke stap in de analyse een rigoureus wiskundig bewijs op te stellen dat alle mogelijke uitzonderingen en randgevallen omvat. De klemtoon ligt bij sommige berekeningen eerder op een pragmatische aanpak, aangezien we hoofdzakelijk geïnteresseerd zijn in het afleiden van formules en rekenalgoritmen en –procedures, die (hopelijk) efficiënt, voldoende accuraat, en gemakkelijk te implementeren zijn, en dus een nuttig hulpmiddel vormen bij het inschatten van de performantie van dergelijke wachtlijnmodellen voor buffers in communicatienetwerken.

Summary

This monograph provides an overview of part of my research with respect to the modelling and analysis of queuing systems that I have carried out since my early years at the SMACS research group. A queuing system can conceptually be regarded upon as any logical or physical entity where customers enter, have to wait for a certain amount of time before being served, after which they leave the system. We encounter queuing systems on a daily basis in our neighbourhood; just think of the waiting line at the bakery shop or supermarket check-out, waiting in the car until a traffic jam is resolved, ··· . The (analytical) study of this kind of systems is usually carried out by means of a suitable mathematical/stochastic model, that describes the way by which customers enter the system (the *arrival process*), the length of the customer service times and the number of servers, the storage capacity of the queue (*system size*), and the order in which customers are served.

The queuing models that are considered here stem from, and were analysed in the context of, the performance assessment of buffers that appear in certain components of telecommunication networks, such as ATM switching elements, buffers in packet-based (such as IP) networks, and so on. In this kind of environment, the 'customers' are data packets that are being transmitted through the network in digital form, and buffers are implemented in the network nodes for the purpose of the temporary storage of those data packets that, for whatever reason – but mostly due to contention with other packets that are vying for the available bandwidth – can not be immediately forwarded to the next node. Due to the digital, and hence discrete, nature of these data packets, these buffer systems are often described by means of a *discrete-time* queuing model. In this setting, time is not a continuous quantity, but a discrete one that follows from the division of the time axis into intervals of equal length, called *slots*, and the discrete nature of the time-parameter is captured by the sequence number of subsequent slots. The service time, or size, of a packet then represents the length of its *transmission time* (expressed in slots), and the servers are the output links, or output channels, in the network node buffer.

The quantities that are of interest in the analysis of these buffer models pertain to the number of packets in the buffer (the *buffer content*), the time spent in the buffer by a packet (the packet *sojourn time*), and the *loss process* of those packets that are being rejected because the buffer is already full when a new packet arrives. Since both the arrival process and the packet size are described by means of a discrete stochastic process, implying that they can be characterised by means of one or more discrete *random variable*(*s*) (drv(s)) and the associated *probability mass function*(*s*) (pmf(s)), the aforementioned quantities are discrete random variables as well, implying that we are primarily interested in performance indices such as

their mean value and variance, (the asymptotic behaviour of) their pmf, ··· . As far as the packet sojourn times are concerned, we will mainly focus on the situation where packets join the queue, and are served, in the same order by which they arrive. The ultimate goal of this kind of research is to derive the necessary formulae that allow us to investigate the impact of the system parameters on these quantities in the system model, and based on that, to establish proper dimensioning rules that are (hopefully) useful for the actual implementation of such a system. Our solution method of preference makes intensive use of so-called *probability generating functions* (pgfs) to represent the relevant random variables and the equations that relate them. In the course of this text it will become apparent that such an approach requires a quite profound knowledge of the related mathematical and stochastic analysis techniques, the main advantage being that the results that are thus derived can be expressed, to the fullest possible extent, in terms of the system parameters. Therefore, this solution technique is well-suited to realise the goals that were formulated above.

After an introductory discussion in Chapter I of some notions and assumptions that will be adopted in our analyses presented throughout this work, and that have been briefly touched upon in the foregoing remarks, we focus our attention on a relatively simple queuing model in Chapter II, where packet arrivals are generated by N (identical) sources, and where the packet arrival process is described by means of a series of random variables that are assumed to be *independent* and *identically distributed*, implying that a single drv suffices to describe the entire packet arrival process. The packet size is set equal to 1 slot, while the number of output channels is represented by c ($\geq$1). At first we consider an infinite-capacity buffer, which induces a number of simplifications in our analysis. For this model we, consider the buffer content and packet sojourn time, which are studied by means of the pgf-based analysis technique mentioned above, which ultimately leads to closed-form expressions for the mean and variance of these quantities, and the asymptotic tail behaviour of their pmf (also called 'tail distribution') based on the 'dominant-pole' approximation technique. In fact, in this part of this chapter, the foundations are laid for some of the computational algorithms and solution techniques that will prove to be useful in the remainder of the work. In the next step, we also consider the case of a finite-capacity buffer, for which we study the loss process in considerable detail. Specifically, we manage to express the main performance measures that pertain to this loss process in a finite-capacity buffer, in terms of the performance indices related to an infinite-capacity buffer, which results in efficient computational algorithms.

The assumption that packet arrivals during subsequent slots are generated by an independent process is not always very realistic. Therefore, in Chapter III we consider a general framework where dependence – or correlation – between the numbers of packet arrivals in contiguous slots can be imbedded in the arrival model. This is done by means of a *discrete-batch Markovian arrival process* (D-BMAP), where an underlying Markov chain

with L states determines the number of packet arrivals in a slot, depending on the state of the Markov chain in the current and preceding slot. Such an arrival process incorporates a considerable degree of freedom to describe dependence between the numbers of arrivals in successive slots, and has been repeatedly applied to model 'realistic' data traffic, such as packetised video streams or telephone conversations. In case of N identical (*homogeneous*) sources, we again derive expressions for the mean and variance of the buffer content. Moreover, we develop an efficient algorithm for computing the tail distribution of this drv that takes multiple poles of its pgf into account. Also, we will establish a relationship between the (characteristics of) the buffer content on the one hand and the packet sojourn time on the other hand, which is not limited to the specific case of a D-BMAP, and is therefore generally applicable. We will illustrate these results by means of a considerable number of numerical examples that enable us to gain some insight in the underlying mechanisms that govern the system's behaviour. Our results will for instance show that the 'burst factor', which is a characteristic of the packet arrival process, to a large extent determines the system performance, at least as far as the moments of the buffer content and the packet sojourn time are concerned. These results are subsequently extended to the case of non-identical (*heterogeneous*) sources. We thereby focus on two distinct ways by which packets that arrive during the same slot are placed in the queue : on the one hand, all arriving packets can be treated in an equal manner, implying that they join the queue in random, or *arbitrary order* (AO), or, alternatively, they can be ordered based on the packet class they belong to (*fixed-order-by-class*, FOC). Also for this scenario, we manage to establish expressions that allow us to compute the main performance indices that are related to the buffer content and packet sojourn time, although we need to mention that for the AO scheme, significant additional numerical computations are mandatory, compared to the FOC scheme. Finally, this chapter is concluded with a case study, where we consider a network node that is fed by packetised telephone traffic under different scenarios that are evaluated in considerable detail. The purpose is to illustrate the efficacy of the results and formulae that were established in the course of this chapter.

Finally, in Chapter IV, we extend the model and results of the previous chapter to the case where the packet sizes are no longer constant and equal to a single slot, but are represented by a general distribution. We therefore confine ourselves to the single-server case $c=1$, since such models with multiple output channels have not yet been solved in a satisfactory manner, to the best of our knowledge. The analysis of the buffer content, yet again based on an expedient and adequate application of pgfs and their properties, makes use of the *supplementary variable* approach in order to be able to deal with it. Making an extensive use of the analysis techniques that were adopted thus far, we yet again obtain (semi-)analytic expressions for the mean, variance, and tail distribution, of the buffer content and packet sojourn time.

To conclude, we would like to emphasise that, in spite of the somewhat ‘mathematical’ nature of the work that is reported, we will not always make an attempt to justify each and every step in our derivations by a rigid mathematical proof that covers all conceivable cases and exceptions. In some instances we prefer a more ‘pragmatic’ approach, since the main objective is to derive formulae, and computational algorithms and procedures that are (hopefully) efficient, accurate, and easy to implement, and therefore constitute a useful tool in the performance assessment of this type of queuing models for buffers in communication networks.

Table of Content

Frequently used symbols and acronyms

AO : arbitrary ordering;

ATM : asynchronous transfer mode;

bgf(s) : batch generating function(s);

B-ISDN : broadband integrated services digital network;

cdf(s) : cumulative distribution function(s);

ccdf(s) : complementary cumulative distribution function(s);

D-BMAP : discrete-batch Markovian arrival process;

drv(s) : discrete random variable(s);

DT : drop-tail;

FCFS : first-come-first-served;

FOC : fixed-order-by-class;

IBP : interrupted Bernoulli process;

i.i.d. : independent and identically distributed;

hcf : highest common factor (or greatest common divisor);

MMBP : Markov modulated Bernoulli process;

PF : Perron-Frobenius;

pgf(s) : probability generating functions(s);

pgm : probability generating matrix;

pmf(s) : probability mass function(s);

stdv : standard deviation;

$a_{n,l}$: the number of sources in state S_l during slot n;

a_l : the number of sources in state S_l during an arbitrary slot;

$\overline{\boldsymbol{a}}$: the $L\times 1$ vector $[a_1 \cdots a_L]^T$ (with joint pgf $A(\cdot)$);

c : the number of channels (or output links) of the buffer's transmission unit;

d_k : the delay of a packet of type k (with pgf $D_k(\cdot)$);

d : the delay of an arbitrary packet (with pgf $D(\cdot)$) ;

$e_n(i)$: the number of packet arrivals during slot n generated by the i-th source;

e_n : the total number of packet arrivals during slot n;

e : the total number of packet arrivals during an arbitrary slot (with pgf $E(\cdot)$);

$\mathbf{E}[\cdot]$: expected value of the argument;

$f_{r,k}$: number of packets of type r that have arrived during the same slot and will be transmitted before a type-k packet;

f : number of packets that have arrived during the same slot and will be transmitted before a randomly chosen packet (with pgf $F(\cdot)$);

$G_{ij}(\cdot)$: bgf, i.e., pgf of the number of packet arrivals if the transition $S_i \rightarrow S_j$ occurs for a D-BMAP source;

$\mathbf{I}$: $L \times L$ identity matrix;

K : number of traffic classes in a heterogeneous packet arrival process;

K_q : storage capacity of the queue;

ℓ_n : the number of packets that are lost during slot n (with pgf $L_n(\cdot)$);

ℓ : the number of packets that are lost during an arbitrary slot (with pgf $L(\cdot)$);

L : the number of states of the underlying Markov chain of the D-BMAP;

$\mathcal{L}$: hcf of a designated set of integers;

$\overline{\boldsymbol{l}}$ and $\overline{\boldsymbol{m}}$: short-hand notation for the $L \times 1$ vectors of integers $[l_1 \cdots l_L]^T$ and $[m_1 \cdots m_L]^T$;

$\mathcal{M}[G(z)]$: operator that returns the mean value of the drv that is associated with the pgf $G(z)$;

$P_n(\cdot)$: joint pgf of the state vector at the start of slot n;

$P(\cdot)$: joint steady-state pgf of the state vector;

P_{LP} : packet loss probability;

P_{LR} : packet loss ratio;

$p(\cdot)$: boundary probability;

p_{lm} : probability that the state transition $S_l \rightarrow S_m$ occurs at a slot mark;

$\Pr[Y]$: probability that the event Y occurs;

$\mathbf{Q}_k(\cdot)$: $L \times L$ pgm of a D-BMAP of type k;

$\mathbf{Q}(\cdot)$: $L \times L$ pgm of a D-BMAP;

q_n : queue content at the beginning of slot n (with pgf $Q_n(\cdot)$);

q : queue content at the beginning an arbitrary slot (with pgf $Q(\cdot)$);

$\mathcal{R}$: radius of convergence of a pgf or pgm;

s_n : system content at the beginning of slot n (with pgf $S_n(\cdot)$);

s : system content at the beginning an arbitrary slot (with pgf $S(\cdot)$);

$s_c(i)$: cdf of s;

t_k : the transmission time of a packet of type k (with pgf $T_k(\cdot)$);

t : the transmission time of an arbitrary packet of type k (with pgf $T(\cdot)$);

T_A : mean length of an active period;

T_P : mean length of a passive period;

T_h : integer threshold;

$\mathbf{U}(\cdot)$: $L\times L$ matrix that contains the right-eigenvectors of $\mathbf{Q}(\cdot)$ (subscript k added if needed);

$\bar{\boldsymbol{u}}_i(\cdot)$: i-th column of $\mathbf{U}(\cdot)$ (subscript k added if needed);

u_n : unfinished work in the buffer at the start of slot n;

u : unfinished work in the buffer at the start of an arbitrary slot (with pgf $U(\cdot)$);

$\mathcal{V}[G(z)]$: operator that returns the variance of the drv that is associated with the pgf $G(z)$;

w_k : the waiting time of a packet of type k (with pgf $W_k(\cdot)$);

w : the waiting time of an arbitrary packet (with pgf $W(\cdot)$);

$\mathbf{W}(\cdot)$: $L\times L$ matrix that contains the left-eigenvectors of $\mathbf{Q}(\cdot)$ (subscript k added if needed);

$\bar{\boldsymbol{w}}_i(\cdot)$: i- th row of $\mathbf{W}(\cdot)$ (subscript k added if needed);

$\bar{\boldsymbol{x}}$: short-hand notation for the $L\times 1$ vector of complex arguments $[x_1 \cdots x_L]^T$;

z_i : zero inside the complex unit disk (subscripts added if needed);

z_o : zero outside the complex unit disk (subscripts added if needed);

$\bar{\boldsymbol{0}}$: the $L\times 1$ vector with all entries equal to 0;

$\bar{\boldsymbol{1}}$: the $L\times 1$ vector with all entries equal to 1;

$\bar{\infty}$: the $L\times 1$ vector with all entries equal to $\pm\infty$;

α : activity grade of a source;

κ_b : burst factor;

κ_c : clump factor;

κ_s : skew factor;

$\lambda_1(\cdot)$: PF-eigenvalue of $\mathbf{Q}(\cdot)$ (subscript k added if needed);

$\lambda_i(\cdot)$: i-th eigenvalue of $\mathbf{Q}(\cdot)$ (subscript k added if needed);

μ_X : mean value of the drv X;

$\mu_{3,X}$: third central moment of the drv X;

σ_X : stdv of the drv X;

ρ : (offered) load of the buffer system;

$\overline{\boldsymbol{\pi}}$: stationary vector of the underlying Markov chain of the D-BMAP;

$\Psi_{\overline{\boldsymbol{m}}}(z)$: function that contains the boundary probabilities;

$\Omega_{\overline{\boldsymbol{a}}}$: sample space of $\overline{\boldsymbol{a}}$.

Chapter I

A brief introduction to discrete-time queuing models

I.1 Preface

Queuing theory can be circumscribed as the scientific discipline, part of applied probability theory, which focuses on the study of queuing phenomena that occur in *queuing systems*. A queuing systems, in a broad sense, can be defined as any system where *customers* enter, may have to *wait* before receiving a certain treatment or *service*, after which they may leave, re-enter, or visit a subsequent system. As such, we have to deal with queues on an almost regularly basis in every-day life : just think of customers lining up in the waiting room of the hospital or at the counter of a retail shop; traffic congestion at a highway, turnpike, train station or airport; delays in the delivery of our electronic or snail mail,…; this explains why queuing theory is at times referred to as the 'mathematics of discontent', as illustrated by [131], [202], [233].

As a scientific discipline, the seeds of queuing theory were planted at the start of the previous century, first by A.K. Erlang and later followed by T.O. Engset, who are widely recognised as being the founding fathers of this field (see also [175], [247] for some background on this). These first models were developed to study the performance of – at first manually and at a later stage electrically operated – telephone exchanges; Erlang published his first work as early as 1909 with the purpose of investigating the holding times of conversations in such an exchange, and his best-known work, on the Erlang-B formula, dates from 1917 (and was translated to French in 1925 to acclaim worldwide recognition; [157]). This research field has since then thrived and been applied in many disciplines ([256]), as far apart as (the study of) health care and emergency planning ([134], [238]), transportation (car, train and air traffic congestion control, [167], [220]), stock management and production process planning ([138], [242]), machine breakdowns and repairs ([158], [254]), database management and computer networks ([189]), and many others. Due to the relatively recent advent of high-speed (broadband) packet-based data networks, the focus has in the course of the last decades for a large part returned to topics that are related to the design and

management of nowadays communication networks. This is also the context in which the models and analyses that are presented in this work have been developed.

I.2 Buffers in communication networks

Queuing systems in communication networks often emerge under the form of *buffers*, which are electronic components or devices that are implemented for the temporary storage and management of information streams – mostly under the form of digital data packets – if these can not be instantaneously forwarded to their next destination. The quality of the multimedia services that are carried by the present-day high-speed – possibly partly wireless – broadband communication infrastructure is to a large extent determined by the performance of the buffers that are encountered by such a stream in the various network components, such as routers, switching elements, modems, access multiplexers/de-multiplexers, network adapters and interfaces, and so on. The reason that such buffers need to be provided, is the highly irregular manner in which the data source(s) may inject packets into the network nodes, with multiple data streams vying for the often limited network resources (such as the available bandwidth or link rate, the buffer storage capacity, ···). The buffer management and scheduling mechanisms that are implemented in the subsequent network nodes, ideally (are supposed to) provide an orderly and fair treatment and transmission of the packets that accumulate in the buffer, in particular if multiple data streams, with varying requirements in terms of the network delay and information loss that can be tolerated, need to be sent over one or more common output channel(s).

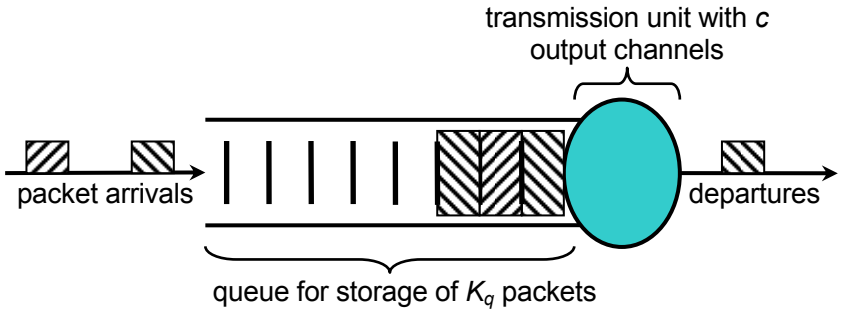


Figure I-1 : buffer model

For the purpose of constructing an adequate mathematical model, such a buffer system, depicted in Fig. I-1, can conceptually be regarded upon as a 'black box' consisting of a storage entity for arriving packets, referred to as the *waiting line* or *queue* of the buffer, with storage capacity K_q, and a transmission unit where packets are processed and transmitted over the available output channel(s), or output line(s). In the remainder, the constant c will stand for the number of output channels of the transmission unit, which is the number of packets that can be processed simultaneously for transmission. If $c=1$, this will be referred to as a *single-server* buffer model, whereas $c>1$ corresponds to a *multi-server* model. It will be

assumed that the output channels are permanently available, i.e., are not subjected to interruptions or failures. Models that do take these into account, so-called vacation or service-interruption models, can be considered as well; such as [4], [16], [23], [25], [29], [32] and the references therein.

The queue capacity on the other hand, which is the maximum number of packets that can be stored in the queue, will be represented by the parameter K_q. In most of the queuing models that are investigated throughout this work, it will often be so that K_q is set equal to infinity, which allows us to derive (semi-)analytic closed-form expressions for the relevant performance measures, as will become clear in the course of this dissertation. However, this assumption does not prevent us from making predictions for the relevant performance indices, such as the packet loss ratio, in a finite-capacity queue. The two parameters c and K_q are inherently determined by the hardware implementation of the buffer system under consideration, and are therefore considered to be known quantities of the associated model.

In order to build a viable queuing model, we also need to specify the way by which packets (attempt to) enter the queue (the packet *arrival process*), are stored in the queue and forwarded to the transmission unit (*buffer management* and *scheduling*), and the processing time in the transmission unit they each require (a packet's *transmission* or *service time*, also called the *packet length*). The notions and conceptions concerning these issues, that we will commonly rely upon throughout this work, will be concisely presented and discussed next. Due to the tight connection between 'buffers' and 'queues', terms such as 'queuing model' and 'buffer model' will be used indifferently in the remainder.

I.2.1 buffer modelling : assumptions, notations and conventions

There are numerous ways by which a buffer system, as described above, can be translated into a mathematical model, and any attempt to give a full and complete overview of buffer models for communication networks, and the way in which these are solved, would be an exercise in futility; one can only refer to the excellent work of authors such as [136], [152], [154], [195], [222], [228], [237], [241], [246], [248], [249], and many more, in addition to the references they contain. Let us commence by indicating that a basic distinction that can be made, is the way by which the time parameter is regarded. The earlier queuing contributions that were mentioned in the prefatory remarks, treated time, and quantities that are related to it (such as the holding time of a telephone conversation), as a continuous quantity, in accordance to one's natural experience. However, from the 1950's on, in the wake of Lindley's important paper on the relation between the system delays of consecutive customers in a single-server queue ([213], in which time could be a continuous as well as a discrete parameter), the first queuing models that treated time as a discrete entity emerged, and the first paper that

explicitly focuses on discrete-time queuing modelling is believed to be the paper by Meisling ([221]).

In such a setting, the time axis is divided into intervals of equal length, called *slots*, where time corresponds to the serial number that is assigned to each individual slot, and therefore becomes a discrete variable. This turns out to be a useful way to model queues that appear in present-day communication networks, where information is represented in digital form, and therefore is a discrete quantity as well.

The packet arrival process : due to the highly irregular and capricious (i.e., 'uncertain') manner by which packets may present themselves to a buffering system, the arrival process will be modelled as a *random*, or *stochastic process*, meaning that we designate a *discrete random variable* (*drv*), say e_n, to represent the number of packet arrivals during slot n, and a *probability mass function* (*pmf*) $\Pr[e_n=i]$, $i\geq 0$, that is associated with each of these drvs. From the derivations that are presented throughout this work, it will (hopefully) become clear that it is often easier to calculate with, and mathematically manipulate, the so-called *probability generating function* (*pgf*) of a drv – rather than its pmf – which can be considered to be the z-transform of the pmf, and is generally defined as

$$E_n(z) \triangleq \mathbf{E}\left[z^{e_n}\right] = \sum_{i=0}^{\infty} z^i \Pr[e_n = i] \quad .$$

For more information on the generic properties of a drv, and the moments, pmf, and the pgf that are associated with it, we refer to Appendix A.

If the set $\{e_n : n\geq 0\}$ forms a series of statistically independent drvs, then the packet arrival process is called an *independent* process. Moreover, if the pmfs $e_n(i)$ are all intrinsically independent of the time index n, then the drvs e_n are considered to be *identically distributed*. If the packet arrival process is both independent and identically distributed, this is referred to as an *i.i.d.* process. If such is the case, this implies that the entire packet arrival process can be represented by a single pmf $e(i)$ and/or pgf $E(z)$, which then describe the number of packet arrivals during any particular slot. Independent packet arrival processes are considered in Chapter II, where the groundwork is laid for the study of the non-independent scenarios that are treated in Chapters III and IV.

In a packet-based communications network, it is generally so that packets are generated by more than one source, and we will in general use the integer N to represent the total number of sources. In the vast majority of scenarios that are of interest to us, it is reasonable to assume that these sources operate independently of each other. In addition, in some cases there is reason to believe that the sources are indistinguishable (in a 'stochastic' sense), implying that they can be represented by a common stochastic process and parameter set. In such a case, the packet arrival process is a *homogeneous* process. If this is not the case, i.e., if

different sources exhibit divergent stochastic properties and are therefore characterised by different processes – which is often related to the different kind of applications or services that these sources may represent – then the corresponding arrival process is called a *heterogeneous* process. Both the cases of homogeneous and heterogeneous packet arrival scenarios will be taken into consideration in this monograph.

The packet transmission times : the packet transmission time, or packet length, is the amount of time, expressed as an integer number of slots, that is needed to process (i.e., to 'transmit', or 'serve') a packet in the transmission unit of the buffer system. Due to the slotted nature of our system description, it is assumed that packet transmissions are synchronised with respect to the slot boundaries (or slot marks), meaning that a packet transmission can start (and end) at slot marks only. If we therefore denote by the drv t_k, with pmf $\Pr[t_k=i]$ and pgf $T_k(z)$, the transmission time of a type-k packet that enters the buffer,

$$T_k(z) \triangleq \mathbf{E}\left[z^{t_k}\right] = \sum_{i=1}^{\infty} z^i \Pr[t_k = i] \quad ,$$

then the set of drvs $\{t_k : k \geq 0\}$ fully characterise the transmission process of subsequent packets. Evidently, we assume that the transmission of a packet requires at least one slot. In the above formula, the index k could represent the sequence number that is assigned to consecutive packets that enter the buffer, or could imply some other or additional typification, depending on the context. Whichever the case : for a homogeneous packet arrival process, it will be assumed that this set of drvs constitutes a set of i.i.d. drvs, which means that a single pgf $T(z)$ characterises the entire packet transmission process. For a heterogeneous packet arrival process, the assumption that consecutive packet transmission times are independent will be upheld, but they may depend on the traffic class that a packet belongs to, in which case the pgf $T_k(z)$ will be used to describe the lengths of all packets of type k. In Chapters II and III, we focus on single-slot packet transmission times, whereas generally distributed transmission times are considered in Chapter IV.

Buffer management and scheduling : generally speaking, 'buffer management' refers to the extent to which packets of specific types are allowed to enter the buffer, while 'scheduling' refers to the order in which those packets that have entered will be transmitted. Although the two concepts can not always be strictly separated from each other, buffer management schemes will primarily have an impact on performance indices that are related to packet loss, whereas scheduling mainly influences the time that a packet spends in the buffer.

The buffer management scheme that will be considered in this work, when the buffer capacity is finite, is the drop-tail (DT) acceptance policy, whereby an arriving packet is allowed to join the queue as long as it is not completely filled. Although most of the models that are analysed in the following chapters assume an infinite-buffer capacity, we should

mention that many studies under a variety of modelling assumptions (e.g., [43], [124], [125],, [153], [192], [262]) have pointed to the fact that a proper use of the results that are obtained for an infinite-capacity buffer analysis permits an accurate prediction of the finite-capacity performance measures, provided that the buffer management mechanism is DT. Many alternative management schemes for buffers in communication networks have been investigated in the past as well; just to mention a few : the *push-out buffer* (PoB, [139], [177], [216]), *partial buffer sharing* (PBS [196], [204]), *random early detection* (RED, [162]), in addition to a considerable number of variations on these themes (e.g., [15], [142], [143] [166], [199], [212], [267] and the references therein).

As far as the scheduling discipline is concerned, we will focus on the first-come-first-served (FCFS) paradigm, in which case packets join the queue (provided that there is an available spot) in the same order in which they arrive. If multiple packets arrive during a slot, one can assume that they will join the queue in *arbitrary order* (AO), since such packets are considered to be equivalent (i.e., indistinguishable), and no particular ranking can be imposed. We should note that this is not necessarily the case for a heterogeneous packet arrival process, and an alternative could be that packets of different types that arrive during the same slot join the queue in a predetermined, or fixed, order, depending on the traffic class they belong to; this ordering mechanism will be designated *fixed-order-by-class* (FOC). Popular alternative scheduling schemes for buffers in a communications environment, such as *head of line* priority ([26], [27]), *random order of service* (RoS, [200]), *weighted round robin* (WRR [17], [191]), *weighted fair queuing* (WFQ, [234]), *virtual clock* (VC, [269]), *earliest due date* (EDD, [159]), and so on, have been investigated and reported on many occasions, but will not be the focus of this work. An insightful discussion and comparison of various buffer management and scheduling schemes can also be found in [149].

I.2.2 assessment of the buffer behaviour

Some of the definitions and notions that are introduced in this section are illustrated in Fig. I-2.

Once the buffer capacity K_q and the number of output links c, the packet arrival and transmission process, and the buffer management and scheduling strategies have been determined and/or appropriately modelled, one can then proceed to the next step, which is the ascertainment of the buffer performance. The behaviour of any queuing system is studied, in general, by investigating one, some, or all of the following quantities : the *number of packets* (or, closely related to that, the amount of work) that the system contains, the *time that a packet spends* in the system, and the *number of packets that are rejected* by the buffer management scheme under consideration. Since the packet arrival and transmission process are being modelled as stochastic processes, these quantities will be (discrete) stochastic

variables as well, and their analysis usually requires the application of a considerable range of mathematical and probabilistic analysis techniques. However, before being able to do so, we need to establish a timeline concerning the events that may take place during a slot. This involves a set of conventions on the manner and order in which packet arrivals and departures occur :

- we will assume that packet arrivals occur somewhere during the course of a slot; the considerations that follow hereafter will reveal that the precise details of the position of the packet arrival instants within a slot are irrelevant for our analysis, which explains why it suffices to characterise the arrival process by a series of drvs describing the *total* number of packet arrivals during subsequent slots without further information concerning their exact arrival instant;
- packet transmissions however, are synchronised with respect to the slot marks, meaning that they can start (and end) at slot marks only. Hence, a packet that enters the buffer during a slot is first stored in the queue, and its transmission is initiated at the earliest at the next slot mark – provided that a server (i.e., transmission line) is available at such a time instant – in which case it is transferred to the system's transmission unit. Consequently, since it takes at least one slot to transmit a packet, a packet that enters the buffer is perforce still in the buffer at the start of (and during) the next slot. This is sometimes called the *please-wait*, or *departures first* (see also [171] and/or [178]) buffering strategy. The opposite scenario is the so-called *come-right-in* procedure, in which case the transmission of new packet arrivals can commence instantaneously upon arrival (such a scheme of course inherently requires the assumption that packet arrivals occur at slot marks).

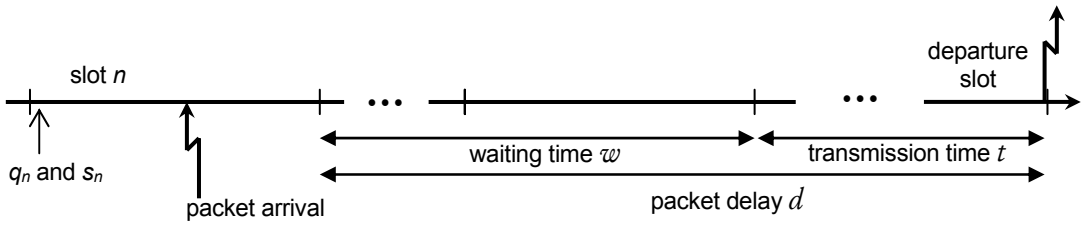


Figure I-2 : quantities of interest when studying the buffer performance

Let us now focus on the amount of packets that are stored in the buffer system. We will denote by the drv s_n the *system content* at the beginning of slot n, which represents the amount of packets in the system, including those that are located in the transmission unit. The main idea is then to express the drv s_{n+1} in terms of s_n by means of a *system equation*, and doing so enables one to study the evolution of the system content at consecutive slot boundaries. For instance, if the buffer capacity is infinite and the transmission time of a packet equals a single slot, then in view of the please-wait buffering strategy we just expounded on (i.e., newly

arriving packets can not be immediately transmitted), then this system equation can be written down as

$$s_{n+1} = (s_n - c)^+ + e_n \quad , \tag{I.1a}$$

where the $(\cdot)^+$ operator signifies $\max\{\cdot,0\}$. If the transmission times are generally distributed, then the system equations become somewhat more complex, as shown in Chapter IV. Similarly, we also define q_n as the *queue content*, which is the number of packets that are lined up in the queue at the beginning of slot n; hence, this quantity does not include the packets that are currently being transmitted, if any. Consequently, the queue and system content will always be related by

$$q_n = (s_n - c)^+ \quad , \tag{I.1b}$$

regardless of the buffer size and packet transmission time distribution. Moreover, in the remainder we will use the term *buffer content* as well, as a generic term for the amount of packets in the buffer, without further specifying whether it either refers to the queue or the system content.

If the queue size K_q is finite, then packets will be rejected as soon as the queue is entirely filled. We mentioned before that we assume that packets are first stored in the queue (during their arrival slot) before being able to move on to the transmission unit. Therefore, since under such conditions the system content at the beginning of a slot is equal to the number of packets that were stored in the queue at the end of the foregoing slot, the drv s_n can not be larger than K_q, and system equation (I.1a), reflecting the case of single-slot transmission times, is now transformed into

$$s_{n+1} = \min\left\{(s_n - c)^+ + e_n, K_q\right\} \quad . \tag{I.2a}$$

In addition, if we represent by ℓ_n the number of packets that are rejected during slot n due to buffer overflow, then under these circumstances this quantity satisfies

$$\ell_n = \left\{(s_n - c)^+ + e_n - K_q\right\}^+ \quad , \tag{I.2b}$$

and the latter two equations fully determine the dynamics of the finite-capacity system under consideration.

Furthermore, the *waiting time* of the k-th packet that enters the system will be represented by the drv w_k, and is defined as the number of slots that the packet spends in the queue, starting from the end of its arrival slot, until the slot mark where it is transferred to the transmission unit. The *delay* of the k-th packet on the other hand, denoted by the drv d_k, is the

total number of slots that a packet spends in the system, from the end of its arrival slot until the end of its transmission time. Clearly, the difference between d_k and w_k is the amount of slots that a packet spends in the transmission unit, and we may write

$$d_k = w_k + t_k \quad . \tag{I.3}$$

In view of the assumptions that were outlined in the previous section, the drvs w_k and t_k are statistically independent in all scenarios that are considered in the next three chapters. We will also use the phrase *sojourn time* as a generic term to refer to the 'time spent in the system', without going into further detail. In our derivations concerning the packet sojourn time, we will rely on the observation that the waiting time of a tagged packet that enters the buffer in case of a FCFS scheduling scheme, is determined by the remaining transmission time of the packet that is being transmitted during its arrival slot, the queue content at the beginning of its arrival slot, and the number of packets that arrive during the same slot and are positioned before it in the queue.

Let us conclude this section by mentioning that for a queuing system with infinite storage capacity, the so-called *equilibrium condition* must be fulfilled if the system is to reach a state of *stochastic equilibrium*, or *steady-state*. The equilibrium condition generally expresses that the average amount of work that enters the system per time unit, must be less than the (maximum) amount of work that can be carried out per time unit. From now on, we will represent by the (offered) load ρ the average amount of work that joins (or attempts to join) the queue per slot and per transmission line, which implies that the equilibrium condition translates into

$$\rho < 1 \quad , \tag{I.4}$$

since the transmission unit can handle c 'units of work' per slot. A queuing system that has reached a state of stochastic equilibrium is characterised by the property that the average amount of work that enters the system (per time unit) is equal to the average amount of work that leaves the system (per time unit). Under these circumstances, irrespective of the initial state one starts from (for instance, an empty system at the beginning of slot 0), quantities such as the queue and system content, the number of packets that are lost per slot, and the packet waiting time and delay, become intrinsically independent of the time index (i.e., the slot and packet serial numbers $_n$ and $_k$) and can be represented by q, s, ℓ, w and d respectively, and deriving the characteristics of these quantities under steady-state conditions is the goal that we aim at.

Under conditions of stochastic equilibrium, one can then invoke one of the fundamental laws of queuing theory, called Little's law ([160], [215]), which for the system that we presently consider expresses that

$$\begin{aligned} \mu_s &= \mu_e \cdot \mu_d \\ \mu_q &= \mu_e \cdot \mu_w \end{aligned}, \tag{I.5}$$

if we let μ_X signify the average value of the drv X. Hence, if one is able to calculate the mean queue or system content, then the mean packet waiting time or delay immediately follow from this result, and vice versa. In the course of the following chapters, these relations will be relied upon as a check on the correctness of our derivations on several occasions.

I.2.3 solution methods

There are various ways by which 'real-life' queuing systems can be studied and analysed, each with their own specific benefits and drawbacks. First, one can choose an *experimental* approach by making measurements on the real system (e.g., a production process in a plant), or a 'test-bed' version of it that is limited in size (e.g., in case of a computer network that geographically covers a large territory), of the quantities of interest. Obviously, this allows one to gather data on a system – and its interactions with the environment – that may be quite complex. On the downside, this is evidently a time-consuming, and therefore expensive, approach, and any change in the system's settings may compel the researcher(s) to recommence the whole process, often without providing much insight in the underlying mechanisms that cause certain phenomena to occur. An alternative can be to construct a computer *simulation* model of the real system, which basically consists of an emulation by means of a computer model, of system's dynamics (for instance based on rules set by the system equations discussed above) that govern its behaviour. This already forces the researcher to make certain 'modelling' choices, for instance concerning the customer arrival and/or departure process in a system node. Still, simulation studies are a lot more flexible than measurements on a real system, and enable one to investigate different and intricate scenarios without too much difficulty. Nevertheless, they can be quite time-consuming, depending on the complexity of the system under investigation, and in general also shed little or no light on the impact of the various system parameters on the perceived behaviour. In addition, simulation results may lack the necessary accuracy concerning the predictions of events that are relatively rare (such as system failures); a drawback that the experimental approach suffers from as well.

In this work, we prefer an *analytical* approach to tackle the various queuing problems that we encounter – sometimes aided by simulation to corroborate our results – which basically boils down to building an adequate 'mathematical' model for the real system, meaning that certain modelling assumptions need to be made, after which the system equations are solved by means of an appropriate analytical method. Hence, this approach often requires finding the right balance between the mathematical tractability of the model on the one hand, and its accuracy on the other hand, and this trade-off is an issue that must always be taken into

consideration when applying any kind of analytically generated theoretic result. Nevertheless, the main advantage of this approach is that it by far provides the best insight into the parameters and/or underlying mechanisms that are responsible for a certain behaviour, and therefore can be a powerful tool in the design and parameter tuning of the 'real' system that the model represents. These solution techniques provide the biggest challenge from a scientific point-of-view, since they require a more-than-casual knowledge of the appropriate mathematical and probabilistic solution techniques (yes, this is actually perceived as an 'advantage' by a researcher in this field).

Several analytic solution techniques have been proposed and used, and their application is often closely connected to the underlying modelling assumptions; consequently, discussing all of them would be all but infeasible within the framework of this short introduction. Let us just mention a few approaches that have gained a lot of popularity within this research field. A widely used method is the *matrix-analytic* solution technique ([129], [203], [227], [228], [263], [274]) which provides a broad and generic framework for the analysis of a wide set of queuing systems. This is based on an intelligent matrix representation of the quantities that are needed in these analyses, and making full use of the properties that these matrices satisfy. One impediment it may nevertheless sometimes suffer from, is the size of the matrices that can be dealt with, which is (roughly speaking) determined by the *state space* of the system. To give some idea, in Section III.6 we will consider some examples for which the number of states that can be visited by the packet arrival process can be 10^6 or larger, and to the best of our knowledge, adopting a matrix-analytic solution technique becomes untenable under such circumstances.

An alternative, and approximate, analytic solution approach is the so-called *fluid-flow* solution technique ([120], [127], [141], [230], [245]), in which case a discrete-time queuing system is modelled as receptacle that is being filled with a fluid of varying intensity, while the output process is determined by the size of the hole in the bottom of the receptacle, thereby neglecting the discrete and packetised nature of the information stream. This approach usually yields a (set of) differential equation(s) that need(s) to be solved, and may actually produce closed-form analytic results in situations where other approaches fail to do so. However, whereas this may yield sufficiently accurate results in some scenarios, its overall accuracy can not always be guaranteed, and difficult to estimate beforehand. This is probably best illustrated by a small example, such as the one depicted in Figs. I-3, where we have plotted the *complementary cumulative distribution function* (ccdf) of the amount of work (expressed in number of slots) – which is the probability that this quantity exceeds a certain threshold T_h – in a infinite-capacity buffer with a single output link of $2 Mb/s$, together with the corresponding result obtained by the fluid-flow approach. Without going into too much detail concerning the packet arrival process, let us just mention that it is based on the case

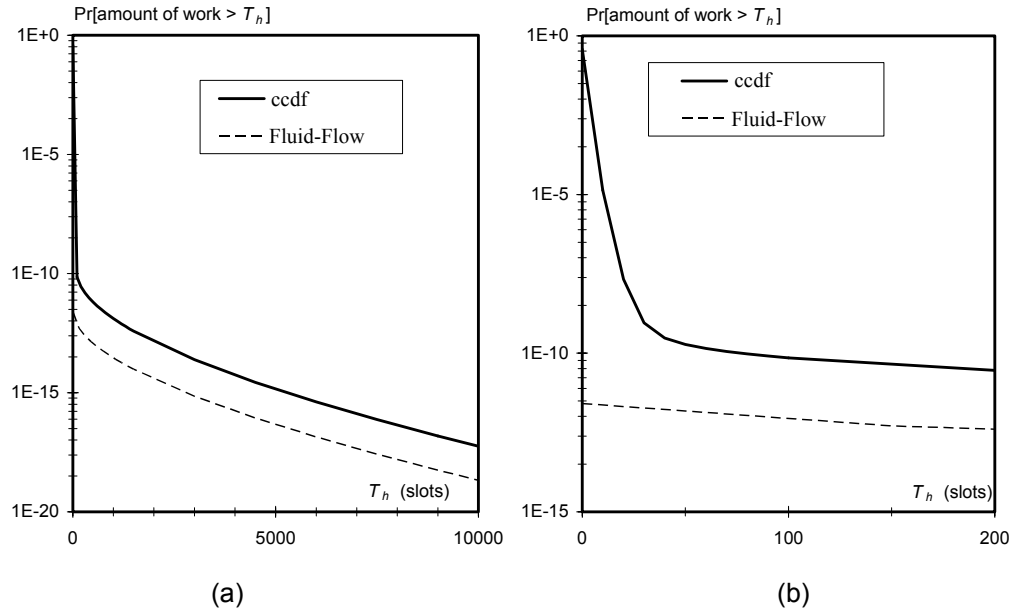


Figure I-3 : ccdf of the amount of work in the buffer on a coarse (a) and fine (b) scale

study presented in Section III.6, of data streams generated by packetised phones that are being multiplexed in a network node. What we wish to highlight is that, although Fig. I-3a reveals that, on a coarse scale for the value of T_h, the fluid flow approach produces a rather satisfactory (albeit too optimistic, since it provides a lower bound for the ccdf) approximation for the slope of the ccdf, these results lack accuracy if we look at a finer scale; e.g. Fig. I-3b. In this respect, it is noteworthy to mention that such a ccdf as depicted here is closely connected to the loss rate of the information stream in a buffer with finite size equal to T_h; consequently, when designing and dimensioning a buffer system, we want to avoid that such a buffer operates in the 'flat' part of the ccdf, where large increases of T_h would induce only marginal reductions of the information loss rate. Hence, these results show that a fluid-flow model may fail to predict where an 'acceptable' point of operation for the system can be found, in which case a more accurate modelling technique becomes mandatory. Also, as far as the 'tail' of the ccdf is concerned, the fluid model underestimates the packet-based model by several orders of magnitude, which may be unacceptable as well. The derivation of computational algorithms for generating such 'packet-based' results, which can be embedded in a basic computer programming procedure, is exactly one of the focus points of this work.

The analytic solution technique that we adopt in Chapters III and IV of this work is related closest to the *spectral decomposition* solution method, which came to prominence by papers such as the ones from S-Q Li, e.g. [209] [210], [211]. This solution technique uses a matrix representation for the quantities of interest, and the main results that are generated still contain operations that can be computationally demanding, such as *Kronecker sums* and *products*, or inversions, of matrices that can have large dimensions. We therefore follow a different path by adopting as far as possible a representation by means of (joint) pgf(s) in our

derivations presented in the three subsequent chapters. The main advantage of this approach is that the formulae for the performance indices that are thereby generated, such as the moments of the queue or system content or the packet waiting time and delay, are – to a large extent – expressed directly in terms of the system parameters, and therefore often lend themselves to an interpretation that is helpful for understanding the mechanisms that determine the buffer behaviour.

I.3 Some background and motivation

As was previously mentioned, most of the work that is presented hereafter has been carried out in the context of ascertaining the performance of buffer systems that arise in various types of data communication networks. An essential component in the nodes of such a network is a *switch*, which, simply put, is a device that gathers data packets from the 'origins' to which it is connected, stores them in (one of) the internal buffer(s) if necessary (and possible), and forwards them to one of its 'destinations'. Hence, the performance of such a switching system is to a large extent determined by the behaviour of the buffer(s) that it contains.

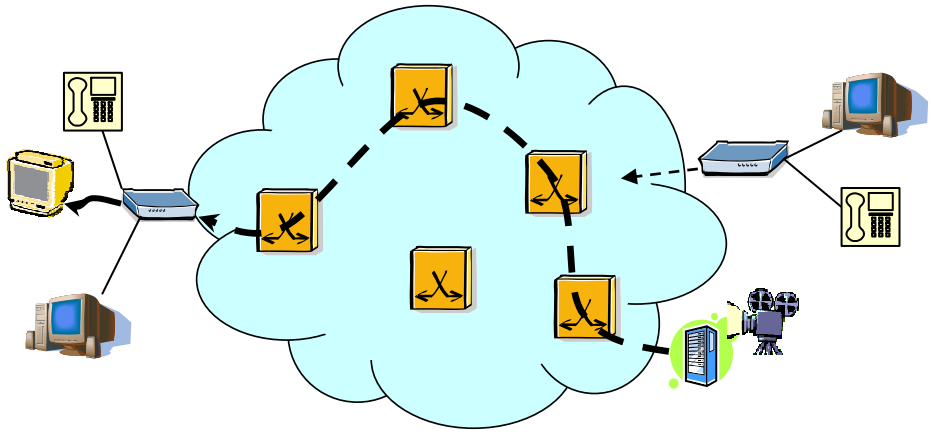


Figure I-4 : a broadband network for integrated services

From the early 1980s on, the *broadband integrated services digital network* (B-ISDN) idea was conceived and developed as a (fixed) networking technology that was a unifying and logical extension of existing circuit-switched – such as telephone – networks. This networking concept aspired to integrate and support all kinds of data services (e.g. Fig. I-4), and fast packet switches for B-ISDN applications have been the focus of many research studies during this time period (e.g. [119], [183], [205], [219], [250], [252], and the references therein). When I started my research in the early 1990s, the ATM (*asynchronous transfer mode*) paradigm was being promoted as the transport technology that would make this ambitious vision come true, and this explains why many of the buffer models that were investigated during that time period and are treated in this monograph, stem from the performance assessment of 'ATM-buffers' or 'ATM-switches' (for some further reading on the ATM transmission paradigm, see [149], [214], [216], [223], [239], among many others).

ATM basically is a *connection-oriented* high-speed packet-switching technology, meaning that a logical connection between the end-users must be established before the exchange of data, under the form of packets, can commence. This data is embedded in ATM *cells* of $53B$(ytes), with a header of $5B$ and a payload of $48B$, which means that adequate segmentation and reassembly algorithms must be implemented at the edges of an ATM network. Once the connection has been established, ATM cells are forwarded in a network node based on the addressing information that they contain, and without too much concern for the kind of data that they carry. Still, in order to deal with the fact that different services may have widely varying requirements concerning the delay and information loss that can be tolerated – the so-called QoS, or quality-of-service – four service categories have been defined : constant bit rate (CBR), variable bit rate (VBR), available bit rate (ABR) and unspecified bit rate (UBR), and the necessary information concerning these traffic types and their specific parameters is exchanged at the connection setup. The divergent service requirements of these four types of traffic are then taken care of by implementing appropriate buffer management and scheduling algorithms in the subsequent ATM nodes that a stream of cells traverses.

One of the primary underlying ideas of the ATM technology is its modular capability, whereby a set of ATM switching *elements* or *modules*, each of relatively small size, can be built into a large-size switching *fabric* that constitutes a node in an ATM network ([205], [155]). ATM switching elements have been proposed in many shapes and forms ([122], [147], [268]) depending on the underlying technology, protocols, and architecture that is implemented, the service classes, buffer management and scheduling schemes that they support, and so on. The one we will presently focus on is the self routing, non-blocking, multipath switching module with output buffering ([180], [214], [235]) that will be briefly expounded upon at the beginning of the next chapter. Many of the buffer models that are reported in this dissertation and that were investigated at the onset of my research (as well as the focus point of many other authors; [136], [174], [182], [190], [231]), in particular those in Chapter II and III, were developed in the context of the performance study of this type of ATM switching module.

From the mid-1990s on however, it became apparent that ATM could not follow through on its promise of being a one-size-fits-all transport technology for a high-speed data network that could support the wide variety of services that were by then finding their way into our homes. Rather, the worldwide data network that has emerged from this era in most cases has IP (*Internet Protocol*) as the dominant protocol implemented at the network layer, which is, to some extent, reflected by the buffer models that are considered in Chapter IV, where packets are no longer treated as entities of a constant length (of $53B$).

Chapter II

Multi-server buffers with independent packet arrivals

II.1 Preface

In this chapter, we analyse the relevant performance measures for buffers with multiple servers and i.i.d. packet arrivals. The inducement for this research was the study of the performance of a symmetric[1] $N\times N(c)$ ATM multipath self-routing switching module with output buffering (e.g., [180], [235]). In this setting, a (data) packet represents an *ATM cell* of $53B$, and therefore has a constant service time, which normally will be chosen such that it corresponds to the length of a single time slot. With the notations and conventions introduced in Chapter I, this implies that $T(z)=z$.

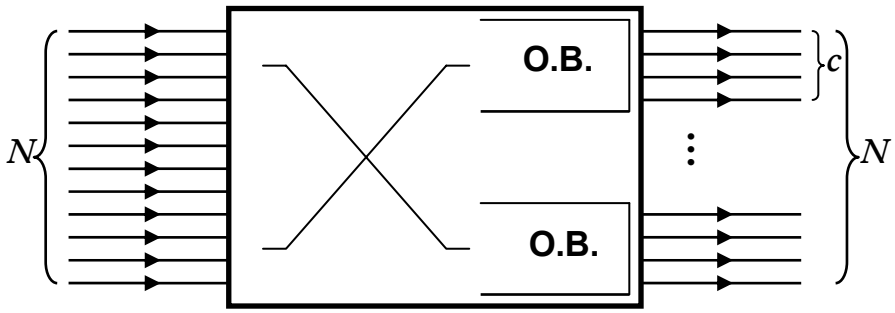


Figure II-1

As illustrated in Fig. II-1, the operation of such an ATM switch can conceptually be summarised as follows. The ingress of the switch consists of N input links via which cells arrive in the switch architecture, while the N output links are grouped into $M=N/c$ clusters or '*output groups'*, each corresponding to a possible 'destination' – which could be an ATM switch at the next stage of an ATM network, an ATM network egress buffer, etc. It is generally assumed that the in- and outlets operate at the same speed. The system parameter c represents the number of output links that is provided per destination, which is also the (maximum) number of cells that can be transmitted per time slot to this particular destination

[1] meaning that the number of input links equals the number of output links of the switch; the situation where this is not the case does not significantly alter the results presented here.

(without loss of generality, N can be considered to be a multiple of c). In short, arriving cells enter the switch architecture, where they are subsequently processed and routed to their respective destinations which can be extracted from the information that is carried by each ATM cell's header that has a fixed length of $5B$. Since the number of cells that arrive during the same slot and that are destined for a particular output group could well exceed c – but is bounded by N – an *output buffer* (O.B.) is provided per destination to anticipate possible contention between these cells. It may in fact be the case that this set of M output buffers is implemented as a single physical shared buffer memory; the M output queues can then be regarded upon as a set of M logical queues, each containing the not-yet transmitted cells that have the corresponding destination in common. Either way, the performance study of such an output buffer is the focus of this chapter.

II.2 The packet arrival process

We will presently confine ourselves to the case where the packet (i.e., ATM cell) arrival process on an input link is an i.i.d. process, with independently distributed destinations as well. This assumption has been made on many occasions when applying queuing models for the performance evaluation of ATM switches; e.g. [147], [153], [172], [174], [182], [183], [190], [200], [217], [231]. Although the derivations in this chapter by no means depend on a specific form of $E(z)$ – the pgf of the drv e that completely captures the i.i.d. packet arrival process in a tagged output buffer – the prefatory remarks concerning the operation of such a switch suggest that the following assumptions can reasonably be made in a first attempt to construct an adequate analytic model for a $N\times N(c)$ ATM switching element :

- the packet arrival process on each of the input links of the switch can be modelled as an i.i.d. *Bernoulli* process with parameter ρ, meaning that the probability that either 0 or 1 packet is generated during any given slot equals $1-\rho$ and ρ respectively, independent of the number of packet arrivals during preceding slots;
- the numbers of packet arrivals on different input links are statistically independent random variables;
- arriving packets are independently and uniformly distributed among the N/c possible destinations, implying that each arriving packet is routed with probability c/N towards a particular output buffer, independent of the decisions that have been made concerning the destination of other packets.

Consequently, the packet arrival process in output buffer j, $1\leq j\leq N/c$, that stems from input link i, $1\leq i\leq N$, is a Bernoulli process as well, with parameter $c\rho/N$. Hence, the overall arrival process in any tagged output buffer during a slot is the aggregation of N identical and independent processes of this type coming from each of the N inlets. Therefore, the numbers

of packet arrivals in output buffer j during consecutive slots constitute an i.i.d. process as well, and the pgf $E(z)$ of the drv e_n that characterises the number of packet arrivals during slot n in the tagged output buffer thereby equals

$$E(z) \triangleq \mathbf{E}\left[z^{e_n}\right] = \left(1+\frac{c\rho}{N}(z-1)\right)^N \quad , \tag{II.1}$$

which corresponds to a binomial packet arrival process with parameters $c\rho/N$ and N respectively, and the mean and variance of the number of packet arrivals per slot satisfy

$$\mu_e = c\rho \quad ; \quad \sigma_e^2 = c\rho\left(1-\frac{c\rho}{N}\right) \quad .$$

The parameter ρ has been defined such that it represents the mean number of packet arrivals per output link, i.e., $\mathbf{E}[e_n]=E'(1)=c\rho$ (adopting standard notational conventions, we let primes ' denote derivatives with respect to the argument).

When N, the number of inlets and outlets of the switch, becomes relatively large – meaning that we let $N\to\infty$ while keeping the aggregate packet arrival rate $c\rho$ constant – then from (II.1) we obtain the following result with respect to $E(z)$:

$$E(z) = \exp\{c\rho(z-1)\} \quad , \tag{II.2}$$

which corresponds to the pgf of a Poisson packet arrival process with mean and variance equal to $c\rho$.

Up to now we have considered a *homogeneous* arrival process (i.e., the packet arrival streams on each of the inlets of the switch are described by identical stochastic processes) with *uniform routing* (i.e., each of the arriving packets has probability c/N of being routed to any of the N/c output buffers). This can of course be easily extended to the case of a (i.i.d.) heterogeneous packet arrival process with non-uniform (but independent) routing. Consider for instance the situation where the packet arrival process on inlet i, $1\le i\le N$, is a Bernoulli process with mean arrival rate γ_i, with $\varepsilon_{i,j}$ the probabilities that an arriving packet on inlet i is routed to output buffer j, $1\le j\le N/c$, that are normalised according to

$$\sum_{j=1}^{M} \varepsilon_{i,j} = 1 \quad .$$

If we let the drv $e_{j,n}$ characterise the i.i.d. packet arrival process in output buffer j during slot n, then its pgf $E_j(z)$ equals

$$E_j(z) = \prod_{i=1}^{N}\left(1+\gamma_i\varepsilon_{i,j}(z-1)\right) \quad , \tag{II.3}$$

and the mean and variance of this arrival process are given by the expressions

$$\mu_{e_j} = \sum_{i=1}^{N} \gamma_i \varepsilon_{i,j} \triangleq c\rho_j$$

$$\sigma^2_{e_j} = c\rho_j - \sum_{i=1}^{N} \left(\gamma_i \varepsilon_{i,j}\right)^2 \quad .$$

Consistent with foregoing definitions for ρ, ρ_j has been defined such that it represents the offered load per output link in output buffer j, i.e., $\mathbf{E}[e_{n,j}]=E_j'(1)=c\rho_j$.

II.3 The buffer content in a multi-server buffer with i.i.d. packet arrivals

The discrete-time queuing model that is being investigated here is by no means new, and an expression for the steady-state pgf of the buffer occupancy in a discrete-time single-server queue with an i.i.d. binomial arrival process and *generally distributed* service times was already reported in [221]. Multi-server models (with single-slot transmission times) are relatively scarce however; e.g. [2], [136], [148], [231].

II.3.1 derivation of $S(z)$ and $Q(z)$

The derivations of this section have been (partly) reported in [2]. Let us consider an output buffer with c output links and infinite storage capacity, depicted in more detail in Fig. II-2. Newly arriving packets are generated according to a general i..i.d. packet arrival process, characterised by the pgf $E(z)$, with mean $E'(1)=c\rho$, for which the binomial and/or Poisson arrival processes, with pgf given by (II.1) and (II.2) respectively, are particular cases.

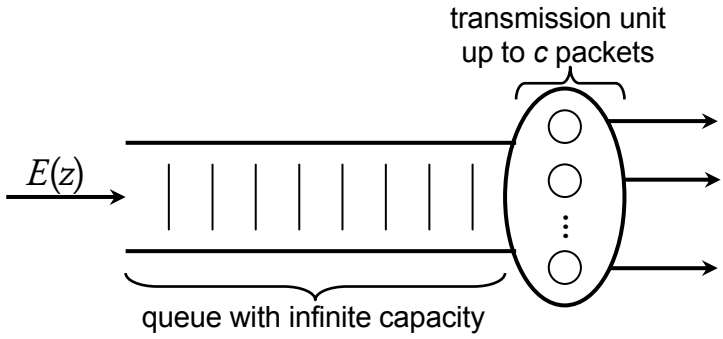


Figure II-2 : an output buffer with c transmission lines and infinite storage capacity

In concurrence with the conventions that were outlined in Chapter I, we let the drv s_n (with pgf $S_n(z)$) represent the system content of this discrete-time queuing system at the n-th slot mark. This is the number of packets that are stored in the buffer at the beginning of slot n, including those that will be transmitted during slot n (if any), and thus leave the system at the end of slot n. Since packets have a transmission time that equals 1 slot, up to c packets will leave the buffer during slot n – provided that as many were already stored in the buffer at the

beginning of this slot. Assuming that the buffer has infinite storage capacity, meaning that no arriving packets will be rejected due to buffer overflow, then the slot-by-slot evolution of the system content can be translated into a *system equation* for the set of drvs $\{s_n : n \geq 0\}$ (see (I.1a))

$$s_{n+1} = (s_n - c)^+ + e_n \quad , \tag{II.4}$$

where $(.)^+ \triangleq \max\{0,.\}$. This system equation expresses that up to c packets that are in the buffer at the beginning of a slot are transmitted during – and leave the buffer at the end of – this slot, while newly arriving packets are stored in the queue and cannot be transmitted yet during their arrival slot. Since by assumption packet arrivals are generated by an i.i.d. stochastic process, the drvs e_n and s_n clearly are statistically independent. Then system equation (II.4) yields the following relation between the corresponding probability generating functions :

$$S_{n+1}(z) = E(z)\mathbf{E}\left[z^{(s_n - c)^+}\right] \quad .$$

The right-hand side of this expression can be rewritten as

$$S_{n+1}(z) = E(z)\left(z^{-c}\mathbf{E}\left[z^{s_n}\{s_n \geq c\}\right] + \Pr[s_n < c]\right) \quad ,$$

where we use the notational convention that, for a function $X(\cdot)$ of one or more drvs and an event Y, $\mathbf{E}[X\{Y\}] \triangleq \mathbf{E}[X \mid Y].\Pr[Y]$. The above equation can then be easily transformed into

$$S_{n+1}(z) = E(z)z^{-c}\left(S_n(z) + \sum_{j=0}^{c-1}\left(z^c - z^j\right)\Pr[s_n = j]\right) \quad . \tag{II.5}$$

We assume that the queuing system under consideration eventually – i.e., for large enough values of n – reaches a stochastic equilibrium, implying that all the drvs that are involved in these derivations become time-independent after a sufficiently long period of time. As a consequence, the drvs s_n and s_{n+1} become equivalent from a stochastic point-of-view, and the associated pgfs $S_n(z)$ and $S_{n+1}(z)$ converge to a common steady-state limit $S(z)$

$$s \triangleq \lim_{n\to\infty} s_n \; ; \; S(z) \triangleq \lim_{n\to\infty} S_n(z) \quad .$$

Thus, $S(z)$ is the pgf associated with the system content at the beginning of an arbitrary slot, which will be denoted by the drv s in the remainder of this work.

The (necessary and sufficient) condition for reaching a stochastic equilibrium can be deduced by expressing that the mean number of packet arrivals that enters the buffer per slot

must be strictly less than the maximum number of packets that can leave the buffer during a slot. This produces the following inequality :

$$c > E'(1) = c\rho \quad \Leftrightarrow \quad \boxed{\rho < 1} \quad , \tag{II.6}$$

which corresponds to the notational conventions outlined in Section I.2.2 (e.g. (I.4)). Under steady-state conditions, the above relation between the pgfs of the system content at the beginning of consecutive slots then yields the following result :

$$S(z) = \frac{E(z)\sum_{j=0}^{c-1}\left(z^c - z^j\right)\Pr[s=j]}{z^c - E(z)} \quad . \tag{II.7}$$

This formula for the steady-state pgf of the system content contains the probabilities $\Pr[s=j]$, $0 \le j \le c\text{-}1$, that are yet to be determined. As indicated in Chapter I, a state of stochastic equilibrium yields the property that the mean number of packet arrivals during an arbitrary slot equals the mean number of packets that depart from the buffer during an arbitrary slot, implying that

$$E'(1) = \sum_{j=0}^{c-1} j\Pr[s=j] + c\sum_{j=c}^{\infty}\Pr[s=j] \;\Rightarrow\; \sum_{j=0}^{c-1}(c-j)\Pr[s=j] \equiv \sum_{j=0}^{c-1} s_c(j) = c(1-\rho) \quad , \tag{II.8}$$

where $s_c(j) \triangleq \Pr[s \le j]$ represents the cumulative distribution function (cdf) of s. Note that this relation comprises the normalisation condition of expression (II.7) for $S(z)$.

The unknown probabilities can be determined as follows. First, we observe that the value $z=1$ is a zero with multiplicity 1 of both the denominator and the numerator of expression (II.7) for the steady-state pgf $S(z)$, and is therefore not a pole of this pgf. Indeed, provided that the equilibrium condition $\rho<1$ holds, and relying on (II.8), it is clear that the first derivative with respect to z for $z=1$ of both numerator and denominator equals $c(1\text{-}\rho)>0$, leading to the conclusion that the multiplicity of the zero $z=1$ of these two functions, is equal to 1. In addition, we can deduce that, by applying Rouché's theorem (e.g. Appendix B**)** on the denominator $z^c\text{-}E(z)$, this function has c zeroes inside the complex unit circle $\{z \in \mathbb{C} : |z| \le 1\}$, among which the zero $z=1$. These zeroes, not including $z=1$, will be designated z_i, $1 \le i \le c-1$.

Now, $S(z)$ being a probability generating function, implies that it is bounded inside the closed complex unit disk (e.g. Appendix A and/or C**)**, and therefore has no poles within this area. Consequently, the z_i's are perforce also zeroes of the numerator of expression (II.7), and this observation imposes a set of c-1 conditions upon the boundary probabilities $\Pr[s=j]$, leading to a set of c-1 linear equations

$$\sum_{j=0}^{c-1}\left(z_i^c - z_i^j\right)\Pr[s=j]=0 \ , \ 1\le i\le c-1 \ ,$$

which, together with (II.8), can be solved numerically. Alternatively, we can point out that the expression in the numerator of (II.7) that contains the unknowns Pr[s=j] is a polynomial of degree c, and the elements of the set $\{z=1\}\cup\{z=z_i : 1\le i\le c\text{-}1\}$ are its c zeroes, as shown above. Hence, this polynomial can also be written as

$$\sum_{j=0}^{c-1}\left(z^c - z^j\right)\Pr[s=j]=(z-1)c(1-\rho)\Psi(z) \ , \ \Psi(z)\triangleq\prod_{i=1}^{c-1}\frac{z-z_i}{1-z_i} \ , \qquad \text{(II.9)}$$

where we have also relied on (II.8) to ensure that the normalisation condition is fulfilled. It is interesting to note that this method of calculating unknown probabilities in the numerator of a pgf by means of the roots with modulus less than 1 of the denominator, has been used as early as [146] in the analysis of the continuous-time $M/D/1$ queuing system. We thus finally obtain

$$\boxed{S(z)=c(1-\rho)\frac{(z-1)E(z)}{z^c-E(z)}\Psi(z)} \ . \qquad \text{(II.10)}$$

This result expresses the steady-state pgf $S(z)$ of the system content in terms of the pgf $E(z)$ describing the i.i.d. packet arrival process, and the zeroes $\{z_i : 1\le i\le c\text{-}1\}$ inside the complex unit circle $\{z\in\mathbb{C} : |z|\le 1\}$. These can be calculated one-by-one, by means of a standard root-finding algorithm such as Newton-Rhapson, by solving the c-1 equations

$$z=E(z)^{1/c}\eta^i \ , \ 1\le i\le c-1 \ ,$$

with $\eta\triangleq\exp\{2\pi\iota/c\}$, the c-th complex root of the value $\exp\{2\pi\iota\}\equiv 1$, and where ι, as usual, represents the imaginary unit in the complex plane[(2)]. Indeed, from Rouché's theorem, it can be deduced as well that each of the above equations has exactly one solution in the region $\{z\in\mathbb{C} : |z|\le 1\}$. The value of the z_i's is plotted in Fig. II-3, for c=32, and a Poisson arrival process with ρ=0.1, 0.3, 0.5, 0.7 and 0.9 respectively. Observe that for $\rho\to 0$, the z_i's converge to η^i, since $E(z)\to 1$ under these conditions.

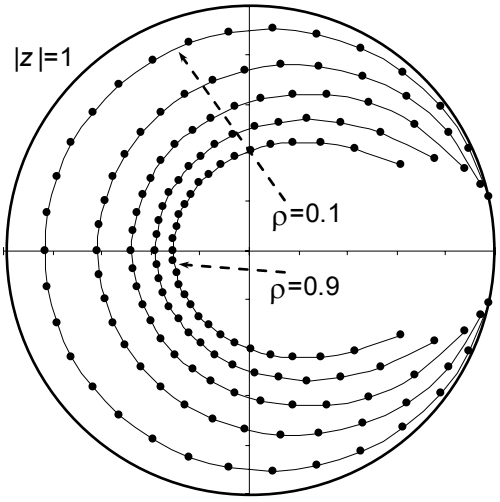


Figure II-3 : value of the z_i's : c=32; Poisson arrivals; ρ=0.1, 0.3, 0.5, 0.7, 0.9

(2) in this work, unless mentioned otherwise, we will consistently refer to 'the n-th root' of a complex number as the *principal branch* of the complex n-th root function, which maps points of the positive real axis into positive real numbers, i.e., if $z\triangleq|z|\exp\{i\theta\}$, then $z^{1/n}=|z|^{1/n}\exp\{i\theta/n\}$.

REMARK

Let $\mathcal{L}$ represent the *highest common factor* (hcf) of the set of integers $\{\{c\} \cup \{n \in \mathbb{N} : \Pr[e=n] \neq 0\}\}$. In other words, the form of $E(z)$ is such that during each slot, packets arrive in multiples of (at least) $\mathcal{L}$. In most situations of interest we will have that $\mathcal{L}=1$, in which case the value $z=1$ will be the only zero of $z^c - E(z)$ with modulus equal to 1 (see also the concluding remarks of Appendix C).

Elaborating on the foregoing analysis, we can also derive an expression for the steady-state pgf $Q(z)$ of the queue content at the beginning of an arbitrary slot, defined as the number of packets in the buffer at this time instant, exclusive of those in the service unit that are being transmitted during this slot. The relation between the system and queue content at the beginning of slot n is obviously given by (as in (I.1b))

$$q_n = (s_n - c)^+ \quad . \tag{II.11a}$$

In combination with system equation (II.4), we find the following relation

$$s_{n+1} = q_n + e_n \quad . \tag{II.11b}$$

Once again assuming that the buffer system reaches a stochastic equilibrium – whereby q (with pgf $Q(z)$) represents the queue content at the beginning of an arbitrary slot in such a system – then invoking the statistical independence between this drv and the number of packet arrivals during this slot, produces the following result :

$$S(z) = E(z)Q(z) \quad . \tag{II.12}$$

In view of expression (II.10), we finally obtain

$$\boxed{Q(z) = c(1-\rho)\frac{(z-1)}{z^c - E(z)}\Psi(z)} \quad . \tag{II.13}$$

This formula expresses the pgf $Q(z)$ in terms of the pgf $E(z)$ that represents the i.i.d. arrival process, and the polynomial $\Psi(z)$ that is completely determined by the c-1 zeroes inside the complex unit disk of the denominator in the right-hand side.

Up to now we have deduced expressions for the steady-state pgfs of the system and queue content, under the condition that the buffer system reaches a stationary equilibrium after a sufficiently long period of time. These results now allow us to derive closed-form formulae for the main performance measures of interest that are related to these quantities, such as their mean, variance, and tail distribution.

II.3.2 the moments of the buffer content

II.3.2.1 mean and variance

Let us commence by defining the differential operators $\mathcal{M}[f(z)]$ and $\mathcal{V}[f(z)]$ of a function $f(z)$, that is assumed to be analytic in z=1, as

$$\begin{aligned}\mathcal{M}[f(z)] &= \frac{d}{dz} f(z)\Big|_{z=1} \equiv f'(1) \\ \mathcal{V}[f(z)] &= \left(\frac{d^2}{dz^2} f(z) + \frac{d}{dz} f(z) - \left(\frac{d}{dz} f(z)\right)^2\right)\Big|_{z=1} \equiv f'(1) + f''(1) - f'(1)^2\end{aligned} \quad , \tag{II.14}$$

When appropriate, we will use $\mathcal{M}[f]$ and $\mathcal{V}[f]$ as the shorthand notation for $\mathcal{M}[f(z)]$ and $\mathcal{V}[f(z)]$ respectively in the following set of equations. Note that whenever $f(z)$ is a pgf, then these two operators reproduce the mean and variance respectively of the corresponding drv. Also, consider $f_1(z)$ and $f_2(z)$, both analytic in z=1; then in addition to the straightforward relations $\mathcal{M}[f_1 + f_2] = \mathcal{M}[f_1] + \mathcal{M}[f_2]$ and $\mathcal{M}[f_1 \cdot f_2] = f_2 \cdot \mathcal{M}[f_1] + f_1 \cdot \mathcal{M}[f_2]$, one can easily verify that (with $f_1 \equiv f_1(1), f_2 \equiv f_2(1)$)

$$\begin{aligned}\mathcal{V}[f_1 + f_2] &= \mathcal{V}[f_1] + \mathcal{V}[f_2] - 2\mathcal{M}[f_1] \cdot \mathcal{M}[f_2] \\ &= \mathcal{V}[f_1] + \mathcal{V}[f_2] + \mathcal{M}[f_1]^2 + \mathcal{M}[f_2]^2 - \mathcal{M}[f_1 + f_2]^2 \\ \mathcal{V}[f_1 \cdot f_2] &= f_2 \cdot \mathcal{V}[f_1] + f_1 \cdot \mathcal{V}[f_2] + f_2 \cdot (1 - f_2) \cdot \mathcal{M}[f_1]^2 + f_1 \cdot (1 - f_1) \cdot \mathcal{M}[f_2]^2 \\ &\quad + 2(1 - f_1 \cdot f_2) \cdot \mathcal{M}[f_1] \cdot \mathcal{M}[f_2] \quad .\end{aligned} \tag{II.15a}$$

If $f_1(1)=f_2(1)=1$, then the following properties are valid :

$$\mathcal{M}[f_1 \cdot f_2] = \mathcal{M}[f_1] + \mathcal{M}[f_2] \; ; \; \mathcal{V}[f_1 \cdot f_2] = \mathcal{V}[f_1] + \mathcal{V}[f_2] \quad , \tag{II.15b}$$

and these identities will be extensively relied upon in this dissertation for the calculation of the first two moments of drvs from their corresponding pgfs.

Let us now consider a general setting where a pgf $G(z)$ of a drv g can be written as the fraction

$$G(z) = \frac{G_n(z)}{G_d(z)} \text{ , where } \begin{cases} G_n(1) = G_d(1) = 0 \\ G_n'(1) = G_d'(1) \neq 0 \end{cases} \quad , \tag{II.16}$$

implying that $G(1)$=1 and that z=1 is a zero with multiplicity 1 of the numerator as well as the denominator in the expression of $G(z)$, a situation that we will encounter on several occasions in this work, such as in the expressions for $S(z)$ and $Q(z)$ that we established in the previous section. With the notational conventions of Appendix A, it can then be deduced – by an expedient use of de l'Hôpital's rule – that

$$\mu_g = \mathcal{M}[G] = \frac{G_n''(1) - G_d''(1)}{2G_d'(1)} \tag{II.17a}$$

$$\sigma_g^2 = \mathcal{V}[G] = \frac{G_n'''(1) - G_d'''(1)}{3G_d'(1)} + \frac{G_n''(1) - G_d''(1)}{2G_d'(1)}\left(1 - \frac{G_n''(1) + G_d''(1)}{2G_d'(1)}\right) \quad , \tag{II.17b}$$

which gives us a procedure to calculate the mean and variance of any drv g whose pgf is of the form (II.16).

If we apply (II.17a) to calculate the expected value of q from expression (II.13) for $Q(z)$, with $G_n(z)=c(1-\rho)(z-1)\Psi(z)$ and $G_d(z)= z^c - E(z)$, we find

$$\mu_q = \frac{E''(1) - c(c-1)}{2c(1-\rho)} + \mathcal{M}[\Psi], \ \mathcal{M}[\Psi] = \Psi'(1) = \sum_{i=1}^{c-1} \frac{1}{1-z_i} \quad , \tag{II.18}$$

which can be (partially) expressed in terms of the first two moments of the i.i.d. packet arrival process as

$$\mu_q = \frac{\sigma_e^2}{2(c-\mu_e)} - \frac{\mu_e}{2} + \frac{1}{2}\sum_{i=1}^{c-1} \frac{1+z_i}{1-z_i} \quad . \tag{II.19}$$

In a similar way, from (II.17b) we can derive the variance of the queue content, leading to the following formula

$$\sigma_q^2 = \frac{E'''(1) - c(c-1)(c-2)}{3c(1-\rho)} + \frac{E''(1) - c(c-1)}{2c(1-\rho)}\left(1 + \frac{E''(1) - c(c-1)}{2c(1-\rho)}\right) + \mathcal{V}[\Psi] \quad ,$$

$$\mathcal{V}[\Psi] = -\sum_{i=1}^{c-1} \frac{z_i}{(1-z_i)^2} \quad . \tag{II.20}$$

If we use the standard shorthand notation $\mu_{3,X}$ for the third (central) moment of the drv X, then an assiduous rearrangement of the terms in (II.20) reveals the following relation between σ_q^2 and the moments of the i.i.d. packet arrival process :

$$\sigma_q^2 = \frac{\mu_{3,e}}{3(c-\mu_e)} + \left(\frac{\sigma_e^2}{2(c-\mu_e)}\right)^2 - \frac{\sigma_e^2}{2} + \frac{1-(c-\mu_e)^2}{12} - \sum_{i=1}^{c-1} \frac{z_i}{(1-z_i)^2} \quad . \tag{II.21}$$

The moments of the system content can be derived from (II.12), by a straightforward application of (II.15b), immediately leading to

$$\begin{aligned} \mu_s &= \mu_e + \mu_q \\ \sigma_s^2 &= \sigma_e^2 + \sigma_q^2 \end{aligned} \quad . \tag{II.22}$$

Expressions (II.19) and (II.21) enable us to calculate the mean and variance of the queue content respectively, as functions of the parameters that determine the packet arrival process, and the zeroes z_i, $1 \le i \le c$-1, that must be calculated numerically. In order to reduce the numerical calculations to an absolute minimum, it would be advantageous to establish accurate approximations for the terms that contain the z_i's in these expressions; this is equivalent to determining accurate bounds for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$.

II.3.2.2 lower- and upper bounds for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$

In view of (II.9), $\Psi(z)$ can be transformed into

$$\Psi(z) = \frac{1}{c(1-\rho)} \sum_{j=0}^{c-1} \frac{z^c - z^j}{z-1} \Pr[s=j] = \frac{1}{c(1-\rho)} \sum_{j=0}^{c-1} z^j s_c(j) \quad , \tag{II.23}$$

where $s_c(j)$ represents the cdf of the system content s. First, observe that, due to this expression for $\Psi(z)$, this function can be regarded upon as a pgf of some drv ψ that takes values between 0 and c-1

$$\Psi(z) = \sum_{j=0}^{c-1} z^j \psi(j) \; ; \; \psi(j) \triangleq \Pr[\psi = j] = \frac{s_c(j)}{c(1-\rho)} \; , \; 0 \le j \le c-1 \quad .$$

Note that $\Psi(1)=1$ due to (II.8). The quantities $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ then equal the mean and variance of ψ respectively, which, evidently, are both positive quantities.

Moreover, upper and lower bounds for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ can be established as well. On the one hand, $\mathcal{M}[\Psi]$ will reach its minimum value when the pmf $\psi(j)$ favours small values of j as much as possible. Bearing in mind that the probabilities $\psi(j)$ are proportional to the cdf probabilities $s_c(j)$, and therefore satisfy the inequalities $\psi(j) \le \psi(j+1)$, $0 \le j \le c$-2, this would be the case for $\psi(j) = \psi(j+1)$, $\forall\, j$,

$$s_c(j) = 1-\rho \;\Rightarrow\; \psi(j) = 1/c, \; 0 \le j \le c-1 \; , \tag{II.24a}$$

implying that ψ is uniformly distributed between 0 and c-1, with mean (c-1)/2, that, in view of the previous remarks, will form a lower bound for $\mathcal{M}[\Psi]$. For the same reason as above (i.e. $\psi(j) \le \psi(j+1)$, $0 \le j \le c$-2), the variance of this uniform distribution, being (c^2-1)/12, will constitute an upper bound for $\mathcal{V}[\Psi]$: no pmf for ψ that satisfies these inequalities and has a higher variance can be (artificially) constructed. Note that $s_c(j)$, $0 \le j \le c$-1, converges to 1-ρ

for $\rho \to 0$, implying that $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ evolve to their respective lower and upper bound under these conditions.

On the other hand, $\mathcal{M}[\Psi]$ will reach its maximum value when the pmf $\psi(j)$ favours high values of j as much as possible. Taking into account the normalisation condition (II.8), and defining $m \triangleq \lfloor c\rho \rfloor$ (where the floor operator $\lfloor \cdot \rfloor$ returns the integer part of the argument), this would be the case for

$$\begin{cases} s_c(j) = 0 & , \ 0 \le j \le m-1 \\ s_c(m) = 1-(c\rho - m) & \\ s_c(j) = 1 & , \ m+1 \le j \le c-1 \end{cases} . \tag{II.24b}$$

The average value of ψ that is associated with these expressions, forms an upper bound for $\mathcal{M}[\Psi]$, and in view of the lower bound derived before, we obtain :

$$\frac{c-1}{2} \le \mathcal{M}[\Psi] \le \frac{(c-m)(c+m-1)}{2c(1-\rho)} - \frac{m(c\rho - m)}{c(1-\rho)} . \tag{II.25a}$$

This also yields the following bounds for the sum that appears in expression (II.19) for μ_q :

$$0 \le \frac{1}{2}\sum_{i=1}^{c-1} \frac{1+z_i}{1-z_i} \le \frac{(c-1)c\rho + m(m+1-2c\rho)}{2c(1-\rho)} . \tag{II.25b}$$

The combination of these inequalities with (II.19) and (II.22) thus provide us with upper and lower bounds for the mean queue and system content, that are easy to evaluate, since they do not require the computation of the z_i's, $1 \le i \le c-1$.

Moreover, the formulae for $s_c(j)$ that were proposed in (II.24b), also produce a pmf for ψ that is centred as tightly as possible around its mean, and the variance that results from it will therefore be a lower bound for $\mathcal{V}[\Psi]$. Hence, we may write

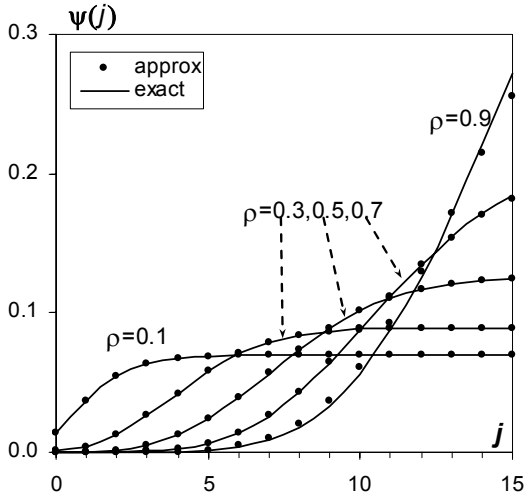


Figure II-4 : pmf of ψ; Poisson arrivals

$$\frac{(c-m)\left(2(c+m)^2 - 2cm - 3(c+m)+1\right)}{6c(1-\rho)} - \frac{m^2(c\rho - m)}{c(1-\rho)} - \mathcal{M}_{ub}[\Psi]^2 \le \mathcal{V}[\Psi] \le \frac{c^2-1}{12} , \tag{II.26}$$

where $\mathcal{M}_{ub}[\Psi]$ represents the upper bound for $\mathcal{M}[\Psi]$ in (II.25a).

The upper and lower bounds for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ given by (II.25a) and (II.26) are sufficiently accurate for the calculation of the mean and variance of the queue and system content, as long as the value of c, the number of servers in the model, is not too high; however, their accuracy somewhat deteriorates for high values of c, as will be shown later on. In order to deal with high values of c, another approach may be followed, that leads to highly accurate approximations for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ regardless of the value of c, and that will also prove to be useful in the analyses of the asymptotic behaviour of the queue and system content in Section II.3.3. This is the reason why we briefly devote some attention to this issue to conclude this section.

The derivation of this approximation is primarily based on the observation that, if the system content at the beginning of a slot satisfies $s \leq j$, then the number of packet arrivals during the previous slot necessarily did not exceed j. Consequently, we can write down the inequality

$$s_c(j) \leq e_c(j) \quad , \quad j \geq 0 \quad ,$$

where $e_c(j) \equiv \Pr[e \leq j]$ represents the cdf of the i.i.d. packet arrival process. Hence, it seems appropriate to introduce the following approximation for $\Psi(z)$

$$\Psi_a(z) \triangleq C_a \sum_{j=0}^{c-1} z^j e_c(j) = C_a \sum_{j=0}^{c-1} \frac{z^c - z^j}{z-1} \Pr[e=j] \; ; \; C_a^{-1} \triangleq \sum_{j=0}^{c-1} (c-j) \Pr[e=j] \quad , \tag{II.27}$$

where the constant C_a ensures the normalisation of this function; all relevant quantities that pertain to this approximation will be marked by the subscript $_a$. The accuracy of this approximation is assessed in Fig. II-4, where we compare the pmf of ψ, as defined before (and calculated as outlined in Section II.3.3.1), with the approximation for this pmf derived from the expression for $\Psi_a(z)$. Considering a Poisson arrival process with c=16 and ρ=0.1, 0.3, 0.5, 0.7 and 0.9 respectively, we can only conclude that the discrepancy between the exact and approximate result is actually quite small. We also observe how this pmf evolves from a quasi-uniform distribution for low values of the offered load, to a pmf that favours high values of i, $0 \leq i \leq c$-1, which supports the approach that was followed to derive upper and lower bounds for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$, e.g. (II.24a,b).

The above remarks immediately lead to the following approximations for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$

$$\begin{aligned} \mathcal{M}[\Psi_a] &= \frac{C_a}{2} \sum_{j=0}^{c-1} (c-j)(c+j-1) \Pr[e=j] \\ \mathcal{V}[\Psi_a] &= \frac{C_a}{6} \sum_{j=0}^{c-1} (c-j)\left(2(c+j)^2 - 2cj - 3c - 3j + 1\right) \Pr[e=j] - \mathcal{M}[\Psi_a]^2 \end{aligned} \tag{II.28}$$

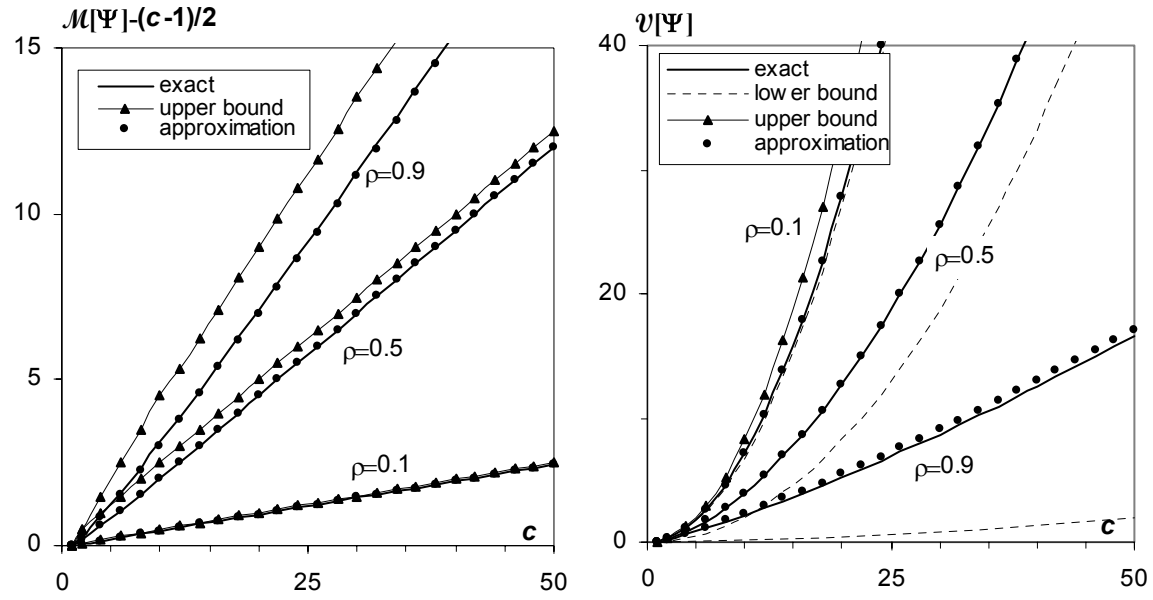


Figure II-5 : approximation and bounds for $\mathcal{M}[\Psi]$ (a) and $\mathcal{V}[\Psi]$ (b)

Due to the presence of C_a in these expressions, it is not possible to determine whether or not $\mathcal{M}[\Psi_a]$ and $\mathcal{V}[\Psi_a]$ form an upper bound for $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ respectively. In view of the results depicted in Fig. II-4 however, we expect them to be very accurate.

In order to get a flavour of the accuracy of the bounds and approximations introduced in this section, we have plotted $\mathcal{M}[\Psi]$-(c-1)/2 and $\mathcal{V}[\Psi]$ and their respective approximation (given by (II.28)) and upper and lower bounds (given by (II.25b) and (II.26) respectively) versus c in Figs II-5a,b. Note that, from (II.25b), the lower bound of $\mathcal{M}[\Psi]$-(c-1)/2 equals 0, while the upper bound (c^2-1)/12 for $\mathcal{V}[\Psi]$ does not depend on the value of ρ. First, we observe that the approximation obtained from (II.28) is highly accurate under all circumstances. Secondly, the upper and lower bounds are quite accurate for low and medium values of ρ, but become less tight as ρ and c increase. However, the contribution of $\mathcal{M}[\Psi]$-(c-1)/2 and $\mathcal{V}[\Psi]$ in the expressions for the mean and variance of the queue and system content becomes less important as ρ increases as well, meaning that equations (II.25a) and (II.26) are still useful to get a first estimate of these quantities when ρ becomes large. If increased accuracy is mandated while at the same time avoiding the calculations of the z_i's, then approximation (II.28) should be invoked.

II.3.2.3 a heavy-load approximation

The queue (or system) content q (or s) in a queuing system with infinite storage capacity typically tends to infinity as the offered load ρ approaches 1 (implying that $\mu_e \rightarrow c$) , and the same holds for its moments, as becomes clear from expressions (II.19) and (II.21) for the

mean and variance of the drv q (or s). If we retain only the most dominant terms in these two expressions, we obtain a so-called *heavy-load approximation* for the mean and variance of the queue and system content

$$\begin{aligned} \mu_{q,\rho\to 1} &= \mu_{s,\rho\to 1} = \frac{\sigma_e^2}{2(c-\mu_e)} \\ \sigma^2_{q,\rho\to 1} &= \sigma^2_{s,\rho\to 1} = \left(\frac{\sigma_e^2}{2(c-\mu_e)}\right)^2 \end{aligned} \quad , \tag{II.29}$$

where we have invoked the formulae (II.25a) and (II.26) for the upper bound of $\mathcal{M}[\Psi]$ and $\mathcal{V}[\Psi]$ respectively, which indicate that the contribution of these terms in the expressions for the mean and variance of the queue content become negligible as $\rho=\mu_e/c\to 1$. Apparently, for fixed values of the load ρ (close to 1), the heavy-load behaviour of both the mean and variance of the queue and buffer content is determined by the variance of the packet arrival process, which can be regarded upon as a measure for the variability of the number of packet arrivals that arrive during a slot.

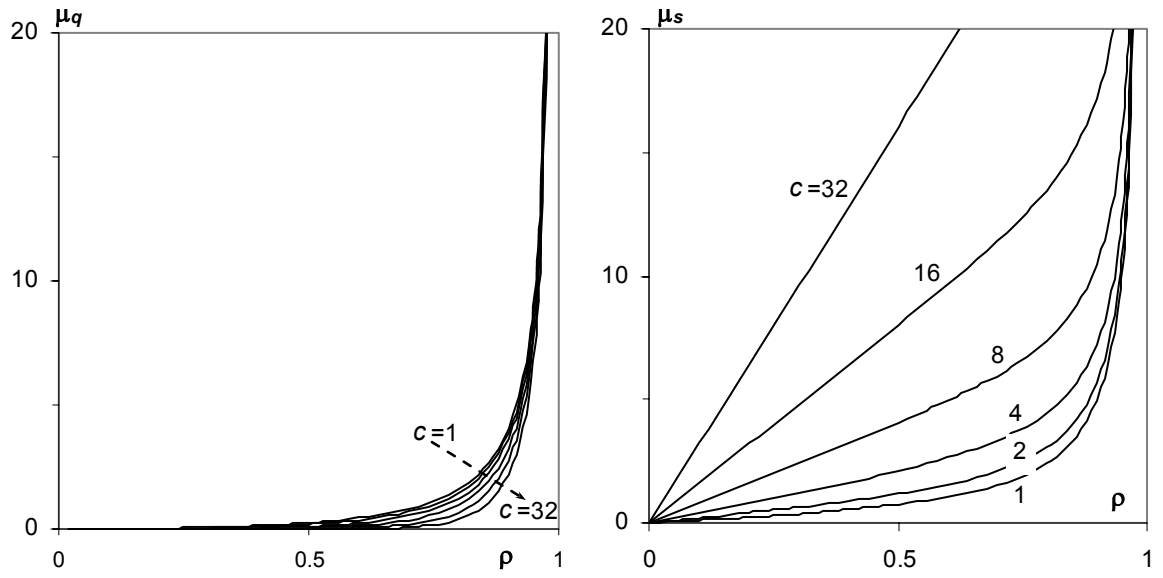
In addition, we observe that the coefficient of variation of the queue and system content, defined as the ratio of their standard deviation versus their average value, tends to 1 as the load ρ becomes close to 1

$$C_{V,\rho\to 1} = \frac{\sigma_{q,\rho\to 1}}{\mu_{q,\rho\to 1}} = \frac{\sigma_{v,\rho\to 1}}{\mu_{v,\rho\to 1}} = 1 \quad ,$$

a conspicuous property that will hold for all queuing models that are considered and investigated in this monograph, as we will demonstrate later on.

II.3.2.4 numerical examples

In Figs. II-6a,b, we have depicted the mean queue and system content versus the load ρ calculated from (II.19) and (II.22), for a Poisson arrival process and c=1,2,4,8,16 and 32 respectively. These plots follow the general behaviour for the mean buffer content in a buffer of infinite size : starting from zero, they steadily increase and reach a vertical asymptote for $\rho\to 1$. We also observe that, while μ_q hardly depends on the value of c, it decreases for increasing c, whereas the opposite is true for μ_s. This points to the fact that, for increasing c, this queuing system increasingly resembles a server-loss system, where the transmission unit is capable of handling most of the incoming packets, and where the queue of the system is, on average, only sporadically used, unless the load is high. This is confirmed by the observation that, especially for large c (e.g. c=32), μ_q is close to zero, whereas μ_s becomes equal to $\mu_e=c\rho$ under these circumstances (remember that $\mu_s=\mu_q+\mu_e$). These assessments are for a large part

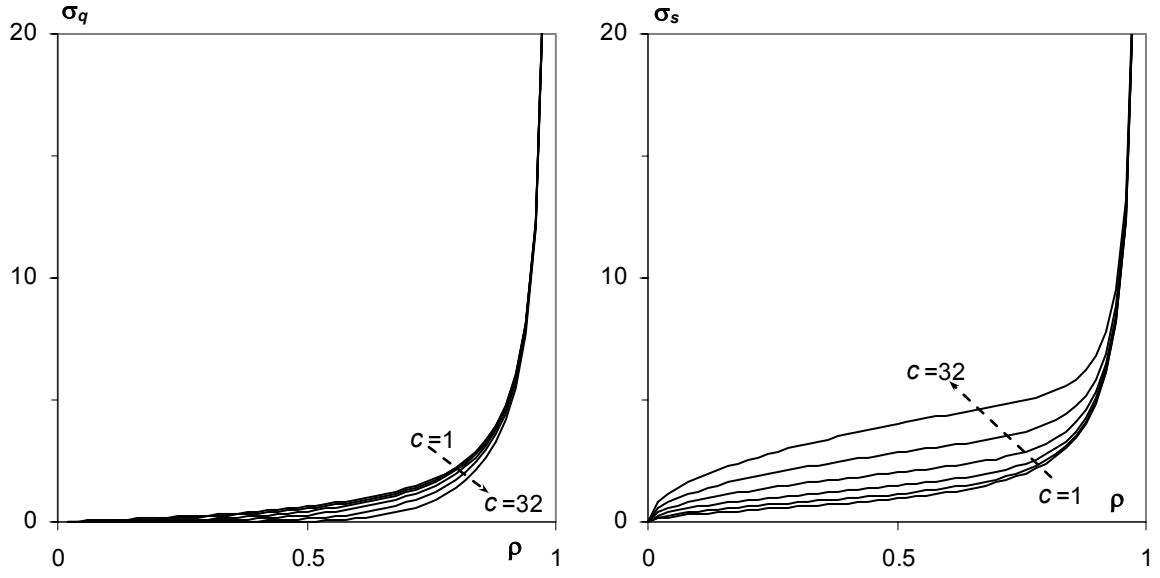


(a) queue content (b) system content

Figure II-6 : mean buffer content versus ρ; Poisson arrivals

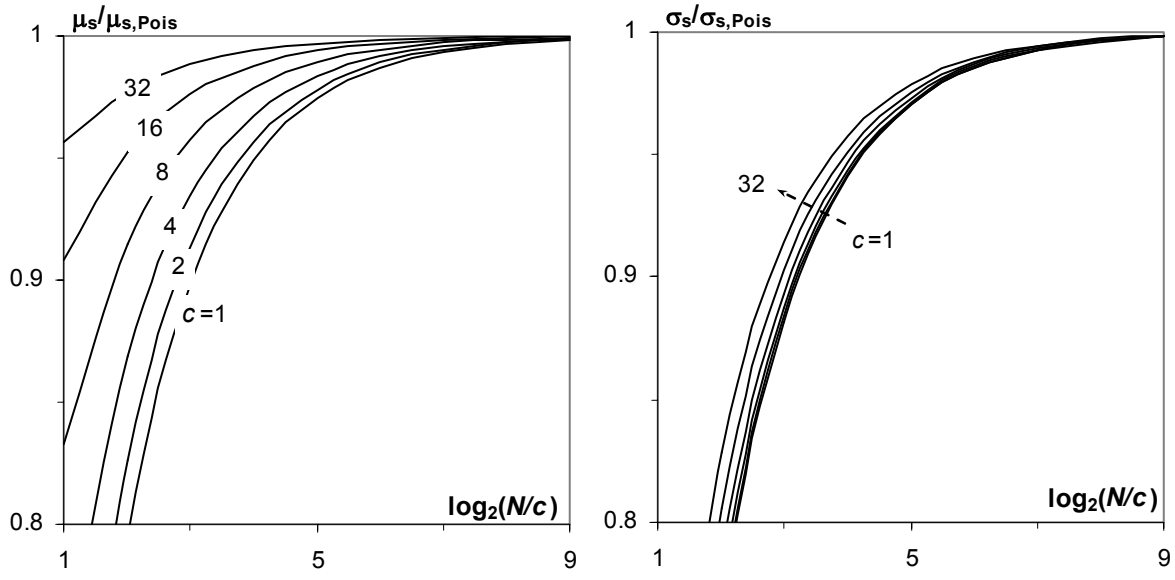
explained by the inherent relative 'smoothness' of the Poisson arrival process under consideration, which is exemplified by the relative low value of its coefficient of variation, equal to $(c\rho)^{-½}$. Apparently, the higher the value of c, the smoother the packet arrival process becomes, and the less likely that more than c packets enter the system during a slot; hence, the less 'queue' we need to anticipate sudden bursts. Similar remarks can be made as far as the standard deviation (stdv) of the queue and system content is concerned, which are plotted in Figs. II-7 for identical values of the system parameters, although this server-loss behaviour (for low and moderate values of ρ) seems to be less pronounced.

In Section II-2, it was reported that the binomial packet arrival process of equation (II.1) becomes a Poisson arrival process for $N\rightarrow\infty$; hence, the moments of the buffer content for a binomial arrival process should converge to the moments of the buffer content in case of a Poisson arrival process as well for increasing values of N, and similar values of c and ρ. This is shown in Figs. II-8a,b, where we have plotted the ratio of the mean (a) and stdv (b) of the system content for a binomial packet arrival process, divided by the mean and stdv of the system content for a Poisson arrival stream, versus $\log_2(N/c)$ (e.g., $\log_2(N/c)=5$ implies that $N=32c$), for a load $\rho=0.9$ and various values of c. These curves indicate that this convergence indeed occurs fairly quickly – for $N/c\geq 32$, the difference between the moments of the system content for the two arrival processes indeed becomes quite small – and with increased speed for higher values of c. The observation that these ratios are less than 1, immediately follows from the property that the variability of a Poisson arrival process, for which the stdv is a relevant measure, exceeds that of a binomial arrival process, and also explains why the former is often used as a worst-case approximation for the latter.



(a) queue content (b) system content

Figure II-7 : stdv of the buffer content versus ρ; Poisson arrivals



(a) mean (b) stdv

Figure II-8 : moments of the system content for binomial arrivals, relative to the Poisson arrivals case

The previous observations indicate that the variability of the packet arrival process will play a crucial role in the overall buffer behaviour. We therefore also focus our attention on a potentially highly capricious arrival process, whose pgf satisfies

$$E(z) = p \cdot \exp\left\{\frac{c\rho}{2p}(z-1)\right\} + (1-p) \cdot \exp\left\{\frac{c\rho}{2(1-p)}(z-1)\right\} \quad , \qquad \text{(II.30a)}$$

which is a weighted sum (with parameter p) of two Poisson processes with mean $c\rho/(2p)$ and $c\rho/2/(1\text{-}p)$ respectively. In addition to $\mu_e=c\rho$, it is not difficult to check that the variance of this packet arrival process is now given by

$$\sigma_e^2 = c\rho\kappa_c \ , \ \kappa_c \triangleq 1+\frac{(1-2p)^2}{4p(1-p)}c\rho \ . \tag{II.30b}$$

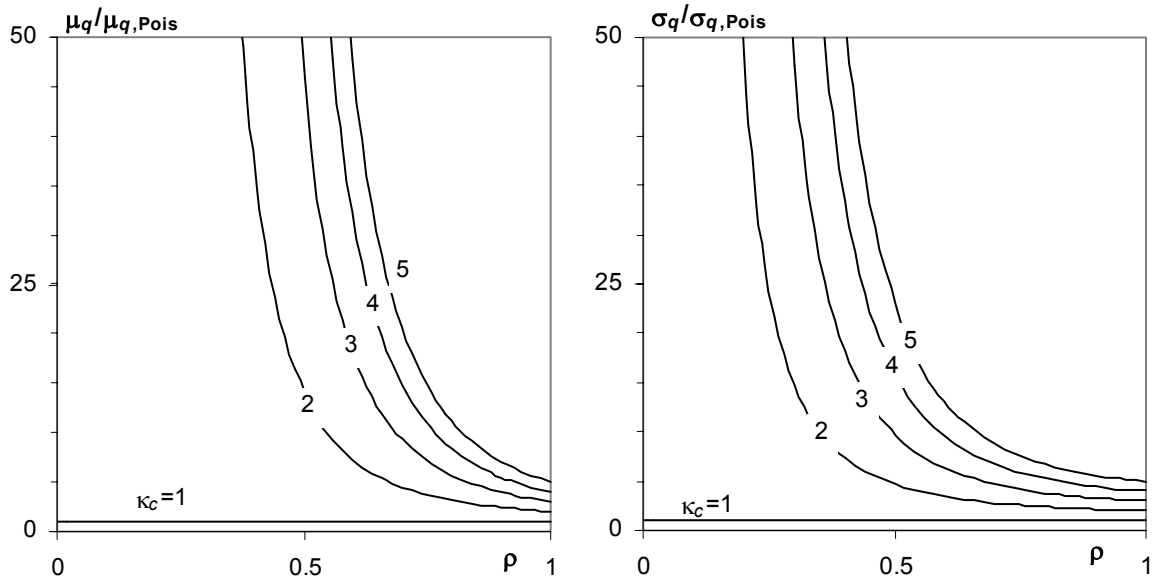
Since $c\rho$ equals the variance of a regular Poisson arrival process with mean $c\rho$, the definition of κ_c ensures that this quantity denotes the ratio of the variance of an arrival process characterised by (II.30a) versus the variance of a Poisson packet arrival process. Therefore, this quantity can be interpreted as being a measure for the (per-slot) variability, or the degree by which packet arrivals are 'clumped' – or grouped – during a slot, compared to a regular Poisson process; hence, the term 'clump factor' will be used to refer to κ_c. Inversely, for a given value of κ_c, the weight parameter p can be calculated from

$$p = \frac{1}{2}\left(1 \pm \sqrt{\frac{\kappa_c - 1}{\kappa_c - 1 + c\rho}}\right) \ . \tag{II.30c}$$

For p=0.5, the arrival process represented by (II.30a) becomes a Poisson process; on the other hand, the closer p approaches to 0 (or to 1), the higher the value of κ_c, and the higher the variability of the arrival pattern that stems from it will be. This arrival process (with κ_c>>1) will therefore be referred to as the 'clumped arrival process'. Note that, in terms of the original ATM switch arrival process described in Section II.2, the pgf $E(z)$ given by (II.30a) can be interpreted as the $N\rightarrow\infty$ limit of

$$E(z) = p\left(1+\frac{c\rho}{2Np}(z-1)\right)^N + (1-p)\left(1+\frac{c\rho}{2N(1-p)}(z-1)\right)^N \ ,$$

meaning that, on each of the N inlets, a packet is still generated with probability ρ during any particular slot, while the routing process is now a 2-mode (independent) process that finds itself in mode 1 (2) with probability p (1-p) during any slot, in which case each of the arriving packets is routed to a tagged output queue with probability $c/(2Np)$ ($c/(2N(1\text{-}p))$); obviously, we must require that N is high enough such that $N \geq c/(2.\min\{p,1\text{-}p\})$ to have a viable routing process. Consequently, although this is still a process that generates packets in an *independent* way (from slot-to-slot) under this interpretation, the routing process no longer treats all packets in an *identical* way, which accounts for the clumping of the packet arrival pattern in the tagged output buffer that results thereof.



(a) mean (b) stdv

Figure II-9 : moments of the queue content for clumped arrivals, relative to the Poisson arrivals case

Figs. II-9a,b show the ratio of the mean (a) and stdv (b) of the queue content relative to the mean and stdv of the queue content in case of a Poisson arrival process, versus ρ, for κ_c=1,2,3,4,5, and c=16. For κ_c=1, the arrival process is a Poisson process, and consequently, $\mu_q/\mu_{q,\text{Pois}}$ and $\sigma_s/\sigma_{s,\text{Pois}}$ are equal to 1. For increasing values κ_c, these ratios drastically increase, due to the strong dependence of μ_q and σ_q on the variance of the number of packet arrivals per slot, e.g. expressions (II.19) and (II.22). We also conclude that, especially for low and moderate values of ρ, these ratios take values that are very high, which points to the fact that the quasi server-loss behaviour that was ascertained in case of a Poisson arrival stream under these conditions, is no longer maintained when the arrival process is increasingly 'clumped', as could be expected. Also, note that the heavy-load formulae (II.29) explain why both $\mu_q/\mu_{q,\text{Pois}}$ and $\sigma_s/\sigma_{s,\text{Pois}}$ approach to κ_c for $\rho\rightarrow 1$.

II.3.3 tail behaviour of the buffer content

The results presented in this section are based on the work that was reported in [1], [6], [47]. In these contributions, as well as in related buffer models – and in particular those that deal with i.i.d. arrival processes, e.g., [124], [262], [165], [257] – it has been observed that adopting a so-called *dominant pole* approximation for the 'tail' of the pmf or ccdf of the buffer occupancy yields remarkably accurate results. The essential features of this approximation technique are discussed next. A more generic approach, that covers a wider variety of queuing models, can be found in [169].

In Appendix A we demonstrate that for a considerable range of discrete-time queuing systems, among which those that are highlighted in this chapter, it is possible to derive an accurate approximation for the asymptotic behaviour of the queue (or system) content from the explicit expression for $Q(z)$ (or $S(z)$), that requires minimal numerical calculations. Let us for now focus on the queue content. This approach essentially boils down to taking the inverse z-transform of $Q(z)$, where only the most significant contribution, namely the one induced by the pole of $Q(z)$ with the smallest modulus, is taken into consideration. Since $Q(z)$, as a pgf, is bounded within the complex unit disk $\{z : |z| \leq 1\}$, then in view of expression (II.13), this pole is the solution of $z^c - E(z) = 0$ outside this area with the smallest modulus. Regarding the existence of such a solution, the following properties hold (for related work, see also [133], [136]) :

Let $\mathcal{R}$ represent the radius of convergence of $E(z)$. As explained in Appendix A, we consider $\mathcal{R} > 1$. Now, consider real values of z, and suppose that condition (C.4)

$$\lim_{\substack{z \to \mathcal{R} \\ <}} E(z) \big/ z^c > 1 \quad ,$$

is satisfied. Note that this also implies that the set of integers $\{n \in \mathbb{N} : n > c \wedge \Pr[e = n] \neq 0\}$ is nonempty, which basically expresses that the queuing system under consideration is non-trivial; otherwise, as explained in Section II.3.2.2, the number of packet arrivals per slot would never exceed c, and we would find that $S(z) \equiv E(z)$ and $Q(z) \equiv 1$. Then, as demonstrated in Appendix C we can prove that

- the equation $z^c - E(z) = 0$ has exactly one positive real solution in the interval $]1, \mathcal{R}[$, denoted by z_o in the remainder;
- z_o is a zero of $z^c - E(z)$ with multiplicity 1;
- the equation $z^c - E(z) = 0$ has no solutions in the region $\{z \in \mathbb{C} : 1 < |z| < z_o\}$.

In addition, let us denominate by $\mathcal{L}$ the *highest common factor* (hcf) of the set of integers $\{\{c\} \cup \{n \in \mathbb{N} : \Pr[e = n] \neq 0\}\}$. Then, provided that $\mathcal{L} = 1$, z_o will in general be the only pole of $Q(z)$ with modulus equal to z_o; if such is not the case, the formulae that were developed in Appendix A can be easily extended to cover this situation. Such a situation is in effect encountered in the packet sojourn time analysis, and how to deal with it is demonstrated in Section II.4.3.

From these properties, it is then clear that z_o is both a simple pole, and the *dominant pole*, of $Q(z)$. By applying the results of Appendix A, see (A.13a), the tail distribution of the drv q can be approximated by

$$\Pr[q=n] \cong \zeta z_{\rm o}^{-n} \quad . \tag{II.31a}$$

an approximation that is very accurate for sufficiently large values of the integer n and that is asymptotically exact, meaning that the exact values for the stationary queue content pmf are found as $n \to \infty$. The constant ζ can be deduced from the residue theorem, which yields (see formula (A.13b) in Appendix A)

$$\zeta \triangleq \lim_{z \to z_{\rm o}} \left(1 - \frac{z}{z_{\rm o}}\right) Q(z) = -\frac{c(1-\rho)(z_{\rm o}-1)}{cz_{\rm o}^{c} - z_{\rm o}E'(z_{\rm o})} \Psi(z_{\rm o}) \quad . \tag{II.31b}$$

By taking the appropriate sum over n in (II.31a), we obtain an approximate expression for the tail of the ccdf of the drv q

$$\Pr[q > T_h] \cong \frac{\zeta z_{\rm o}^{-T_h}}{z_{\rm o}-1} \quad , \tag{II.31c}$$

which allows us to calculate the probability that the steady-state queue content exceeds a given threshold T_h, for large enough values of T_h.

From these results and relation (II.12) (or adopting an analogous approach as outlined in Appendix A), we can deduce similar formulas that capture the asymptotic behaviour of the pmf and ccdf of the system content s, leading to

$$\Pr[s=n] \cong \zeta z_{\rm o}^{-n+c} \quad , \quad \Pr[s > T_h] \cong \frac{\zeta z_{\rm o}^{-T_h+c}}{z_{\rm o}-1} \quad , \tag{II.32}$$

where we have also made use of the property that $E(z_{\rm o}) \equiv z_{\rm o}^{c}$. We would also like to point out that relation (II.11a), under steady-state conditions, implies that

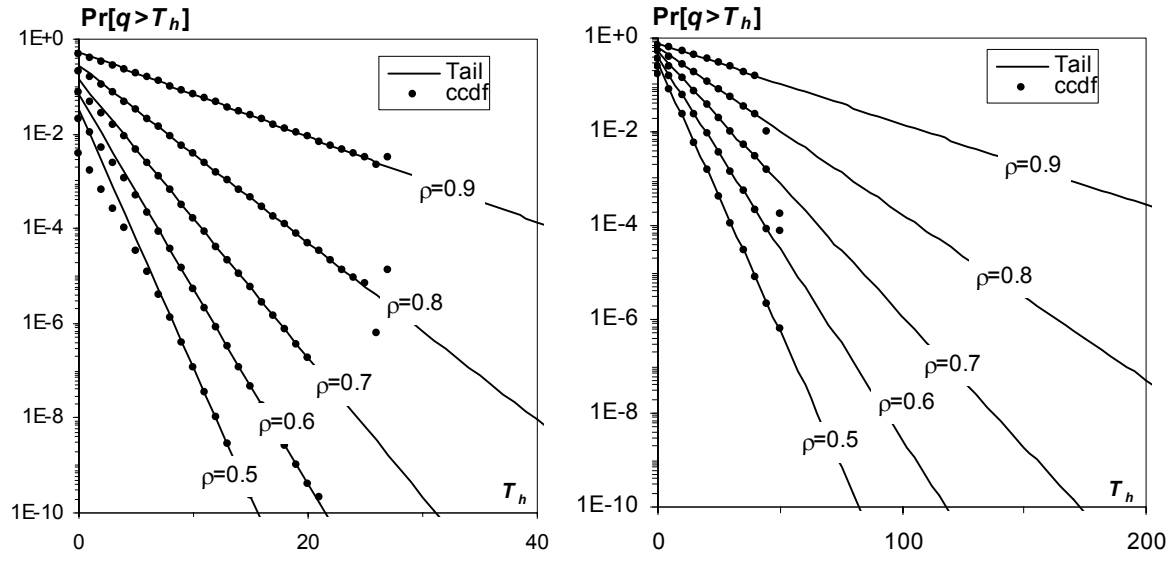
$$\Pr[q > T_h] = \Pr[s > T_h + c] \quad , \; T_h > 0 \quad ,$$

which substantiates the formulae (II.31c) and (II.32).

We conclude by emphasising that the expression for ζ contains the zeroes of z^c–$E(z)$ inside the complex unit disk through the quantity $\Psi(z_{\rm o})$. Once more, the calculation of the z_i's can be avoided by employing approximation $\Psi_a(z_{\rm o})$ for $\Psi(z_{\rm o})$, with $\Psi_a(z_{\rm o})$ given by (II.27).

II.3.3.1 numerical examples

Let us commence by indicating that the cdf $s_c(j)$ of the system content, for $0 \le j \le c$-1, can be derived from (II.23) by identifying the coefficients of z^j in these two expressions for $\Psi(z)$. Also, the system equation (II.4) implies that, for increasing values of j, the $s_c(j)$'s satisfy



(a) κ_c=1 (b) κ_c=5

Figure II-10 : the queue content ccdf; c=16

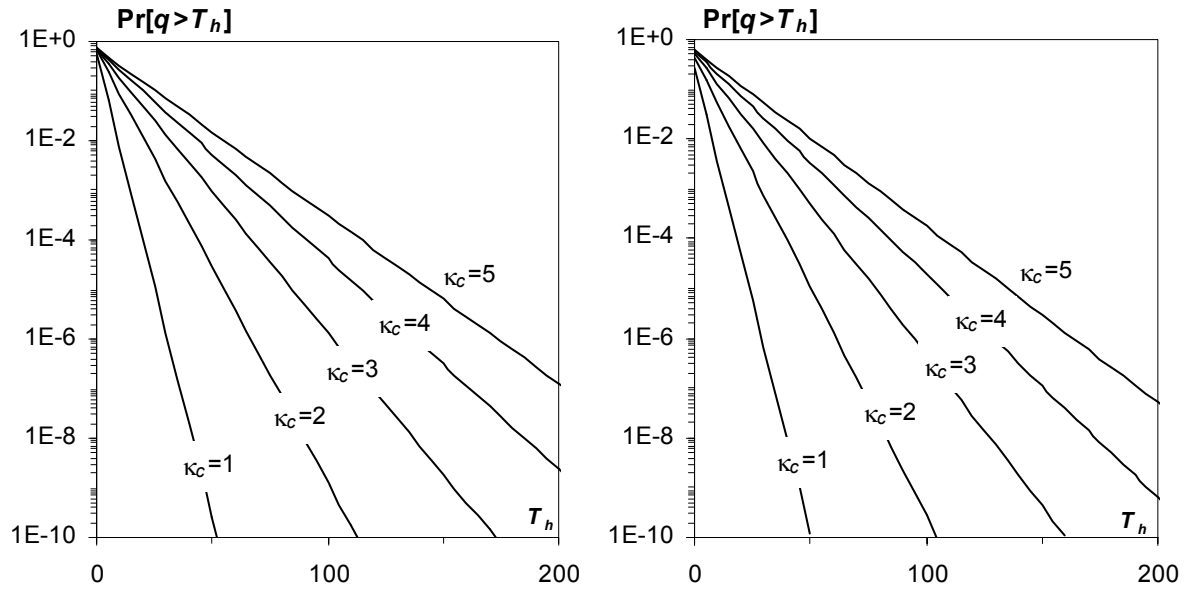
$$s_c(j)=\frac{1}{\Pr[e=0]}\left\{s_c(j-c)-\sum_{i=c}^{j-1}s_c(i)\Pr[e=j-i]\right\}\ ,\ j\geq c\quad , \tag{II.33}$$

and due to the property that $q_c(j)=s_c(j+c)$, $j\geq 0$, this procedure immediately leads to the cdf (and the ccdf 1-$q_c(j)$) of the queue content as well. Due to the close similarities between the ccdf tail approximation of the queue and system content (e.g. (II.31c) and (II.32)), we focus on the former quantity.

In Figs. II-10a,b, we compare the ccdf of the queue content, calculated with the above formula, with its tail approximation, for a Poisson (a) and clumped (b) arrival process with κ_c=5, for $c=16$ and ρ=0.5,0.6,0.7,0.8,0.9 respectively. We observe that, in particular for the higher values of ρ, the ccdf of the queue content quickly converges to its tail approximation, which highlights the accuracy of the approach – of solely taking into account the term that corresponds to the dominant pole – that was expounded in the foregoing section. These figures also illustrate the limited applicability of (II.33), which, for increasing j, rather quickly runs into deadlock due to an accumulation of rounding errors (this is especially true for increasing values of c and κ_c); this in effect points to the efficacy of the procedure for deriving the tail distribution, whose complexity is totally independent of the system parameters[3] and/or the order of magnitude of the probabilities that need to be calculated.

In Section II.5, we will demonstrate that there is a strong correspondence between $\Pr[q>B]$ (or $\Pr[s>B+c]$), and the packet loss probability in a finite buffer of size B; for our

[3] there is a slight dependence on the value of c, since the zeroes z_i, $1\leq i\leq c$-1, need to be calculated; however, this does affect neither the accuracy nor the efficacy of the method



(a) c=1 (b) c=16

Figure II-11 : tail of the the queue content ccdf; ρ=0.8

current purpose, it suffices to state that these probabilities can be interpreted as being a measure for this loss ratio. Then, from Figs. II-10a,b, we can already deduce that the clump factor κ_c will have an undeniable impact on required buffer size if excessive packet loss is to be avoided. This is further illustrated in Figs. II-11a,b, where we have plotted the ccdf tail asymptote of the queue content, for $c = 1$ (a) and 16 (b) respectively, ρ=0.8, and κ_c=1,2,3,4,5. Whereas the value of c has virtually no impact on the depicted tail behaviour, there is a strong dependence on the value of κ_c. Indeed, for instance the intersects with the abscis of these curves suggest a proportionality between the value of κ_c and the required buffer size B for a given loss probability.

The previous observations underscore that in addition to considering $\Pr[q > B]$ as a function of B, the inverse problem, in fact, is an important one as well : for a given target value of $\Pr[q>B]$, what is the value of B that corresponds to it ? We therefore define the 10^{-X} *quantile* of a drv r as

$$Q_r(X) \triangleq \min_x \left\{ \Pr[r > x] < 10^{-X} \right\} \quad , \tag{II.34}$$

i.e., $Q_r(X)$ returns the smallest value of x for which the inequality in the right-hand side still holds. In case that $\Pr[r>x]$ can be approximated by a formula such as (II.31c), then this probability can be interpreted as being a continuous function of x, and under these conditions, the 10^{-X} quantile of the queue content for instance satisfies

$$Q_q(X) = \ln(z_o)^{-1}\left(\ln(\zeta) - \ln(z_o - 1) + X\ln(10)\right) \quad . \tag{II.35a}$$

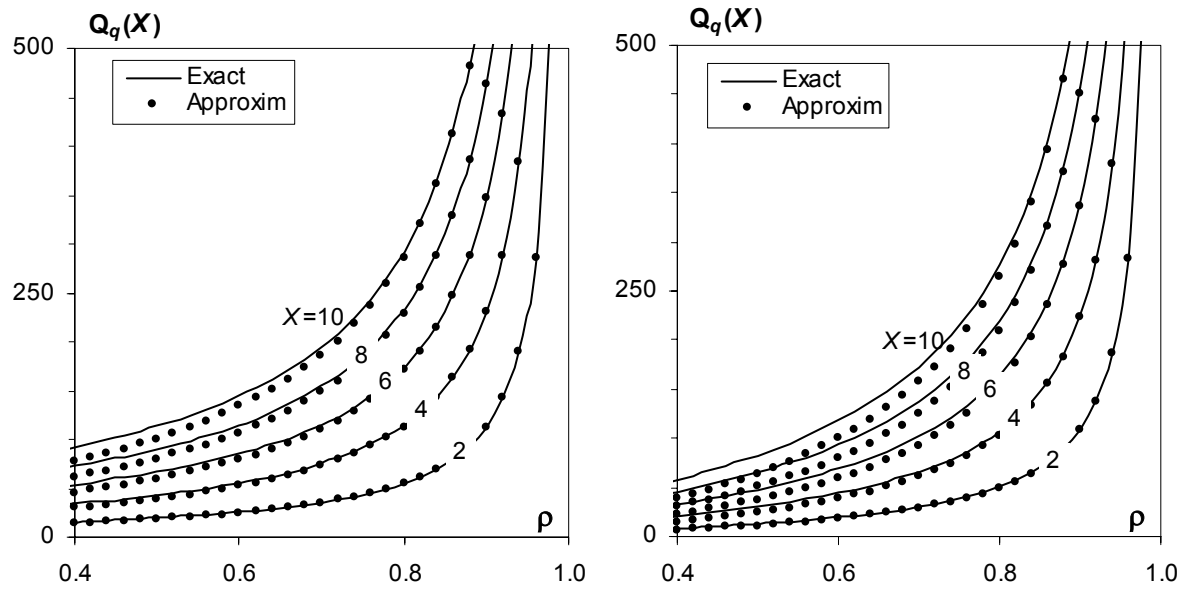


Figure II-12 : quantile of the queue content versus ρ; clumped arrival process with κ_c=5

This may be a good opportunity to mention a rather remarkable (approximate) relationship between the quantile of a random variable, and its first two moments. This is based on some of the results that were reported in [84], where (among others) the quantile of the waiting time in a continuous-time $M/G/1$ queuing system was expressed in terms of its mean and stdv. Translated to the current context, where the buffer system under study can be circumscribed as a discrete-time GI/D/c system, an analogous approach would lead to the following formula that relates the quantile of the queue content with its mean and stdv :

$$Q_q(X) \cong 1 + \mu_q + (X.\ln(10) - 1)\sigma_q \quad , \tag{II.35b}$$

which would imply that the quantile of the queue content is (approximately) proportional to the stdv σ_q, with a proportionality factor that merely depends on the targeted probability 10^{-X}.

In order to assess the accuracy of this formula, we have depicted $Q_q(X)$ – calculated from (II.35a) – versus ρ in Figs. II-12a,b, for a clumped packet arrival process with κ_c=5, and c=1 (a) and 16 (b) respectively, and compared it with the heuristic formula (II.35b). From these results, we can only conclude that there is a striking, if not intriguing, correspondence between the quantile of this drv and the approximation that is suggested above, especially for values of ρ that are not too low. The approximate quantile formula (II.35b) also reveals the reason why this quantity (and hence, the required buffer space in a buffer of finite size) is proportional to κ_c : since both μ_q and σ_q are (roughly, in view of the heavy-load results) proportional to σ_e, for which κ_c is a measure, so is the queue content quantile. This result also suggests that the tail behaviour of the queue content is for the largest part determined by σ_q –

and hence, the variability of the packet arrival process – and also implies the following (approximate) relationship between σ_q and the dominant pole z_o

$$z_o \cong e^{1/\sigma_q} \quad .$$

These are quite interesting issues that undoubtedly merit further examination and refinement, but that will not be pursued here any further, since doing so falls far beyond the scope of this work.

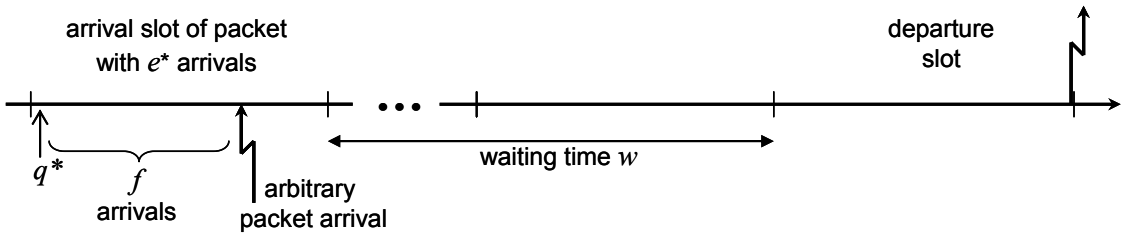


Figure II-13 : the waiting time of an arbitrary packet, with related quantities

II.4 The system sojourn time

In this section, the delay performance of a discrete-time multi-server queuing system with infinite waiting room and general i.i.d. arrivals will be evaluated, by means of a purely analytic technique that is – once more – rooted in the expedient use of probability generating functions. For a first-come-first-served (FCFS) queuing discipline, we derive an expression for the pgf of the system delay and waiting time of an arbitrary packet, focusing on the case that the transmission time of such a packet equals a single slot, as before. This leads to explicit formulas for such performance measures as the mean value, variance, and tail distribution of these quantities. These results are useful, for instance, for studying the performance of an ATM switching element with output queuing, such as described in Section II.1. The results of this section have for the largest part been reported in [2], and are related to [4] as well.

II.4.1 derivation of $W(z)$ and $D(z)$

Let us commence by deriving an expression for the pgf $W(z)$ of the *waiting time* (denoted by the drv w) that an arbitrary packet experiences in the buffer, which is the number of slots spent in the queue, under a first-come-first-served queuing discipline. As depicted in Fig. II-13, the waiting time of a packet is defined as the number of slots between the end of the slot of arrival of the packet, and the beginning of the slot during which its transmission commences. Since the transmission time of a packet equals a single slot, the departure slot of a packet is its transmission slot as well. Note that Fig. II-13 also contains some additional quantities that will prove to be useful during the course of the subsequent analysis. Since the

delay of a packet, represented by the drv d with pgf $D(z)$, is defined as the time that a packet spends in the buffer system, *including* its transmission time, we may obviously write

$$\begin{aligned} d &= w+1 \\ D(z) &= zW(z) \end{aligned} \quad . \tag{II.36}$$

Let us focus on an arbitrary packet, hereafter also referred to as the *tagged* packet that enters the buffer under steady-state conditions during a particular slot. Due to the first-come-first-served nature of the queuing discipline, the waiting time of such a packet is determined by the amount of packets it finds in the buffer upon arrival, which consists on the one hand of the packets that were already in the queue (i.e., the queue content) at the beginning of its arrival slot, and the packets that have arrived during the same slot of the tagged one but 'before' it, on the other hand. Note that the delay and/or waiting time of a packet is by no means influenced by the number of packets that are being transmitted during the tagged packet's arrival slot, which explains why we attempt to express the waiting time in terms of the queue content, rather than the system content, at the beginning of the tagged packet's arrival slot.

Generally speaking, the arrival slot of this packet is not an arbitrary slot, since it is a slot during which there is at least one packet arrival; therefore, we denote by the drv q^* the queue content at the beginning of the tagged packet's arrival slot. However, due to the i.i.d. nature of the packet arrivals process under consideration, the buffer content at the beginning of a slot is independent of the number of packet arrivals thereafter (or the condition that there is at least one such arrival during the subsequent slot). Hence, the drv q^*, and the drv q representing the queue content at the beginning of an arbitrary slot, are equivalent from a stochastic point-of-view, and we may write

$$\mathbf{E}\left[z^{q^*}\right] = Q(z) \quad ,$$

with $Q(z)$ given by expression (II.13).

Furthermore, we define the drv f as the number of packets entering the buffer during the same slot as a tagged packet, but 'before' this packet. This implies that we are somehow able to order all the packets that arrive during the same slot. The standard approach that will be followed in this thesis (unless mentioned otherwise) is to assume that packets that arrive during the same slot are ordered in a random fashion, i.e., they all have equal probability to be positioned first, second, third, etc …, in this ordering process. Since all packets that enter the buffer during a slot receive equal treatment, this random ordering process only depends on the total number of packets that arrive during the tagged packet's arrival slot, which we denote by the drv e^*.

We can determine the pmf of e^* as follows. Consider a series of M consecutive slots, and let n_i represent the number of times that there are i packet arrivals. If we define $r_M(i) \triangleq n_i / M$, and assume that the limit

$$r(i) \triangleq \lim_{M\to\infty} r_M(i) \ , \ \forall i \geq 0 \ ,$$

exists – which is the case for any i.i.d. packet arrival process – then obviously $r(i) \equiv \Pr[e=i]$, $i \geq 0$, i.e., $r(i)$ will be equal to the pmf $e(i) \equiv \Pr[e=i]$ of the i.i.d. packet arrival process. To be more precise, application of Bernoulli's limit theorem ($\forall\ i \geq 0$) on the i.i.d. Bernoulli process that returns the value 1 if the number of arrivals equals i, and 0 otherwise, proves that the series $r_M(i)$ converges *in probability* to $e(i)$, meaning that $\Pr[|r_M(i) - e(i)| > \varepsilon] \to 0$ if $M \to \infty$, for arbitrarily small values of ε. Alternatively, invoking the strong law of large numbers for this i.i.d. Bernoulli process, leads to the conclusion that $r_M(i)$ *almost surely* converges to $e(i)$, i.e., $\Pr[r(i)=e(i)]=1$. Either way, convergence of the relative frequency $r_M(i)$ towards $e(i)$ is assured.

In addition, we observe that among a total of $\sum_j jn_j$ packet arrivals during these M slots, there are in_i packets that arrive during a slot with a total of i arrivals. Hence, if we randomly pick a packet among those that arrive during these M slots (for $M \to \infty$), then the probability that this packet arrives during a slot with a total of i packets can be calculated from

$$\Pr[e^* = i] = \lim_{M\to\infty} \frac{in_i}{\sum_j jn_j} = \lim_{M\to\infty} \frac{in_i/M}{\sum_j jn_j/M} = \frac{ie(i)}{\sum_j je(j)}$$
$$= \frac{ie(i)}{E'(1)} \ , \ i \geq 1 \ ,$$

where, as before, $E'(1)$ represents the mean number of packet arrivals per slot. The existence of these limits is ensured if e has a finite mean and variance.

On the other hand, since our tagged packet is chosen randomly among e^* packets, the conditional pmf of the drv f (with pgf $F(z)$) satisfies

$$\Pr[f = j \,|\, e^* = i] = \frac{1}{i} \ , \ 0 \leq j \leq i-1 \ .$$

Combining the two previous formulae, we thus obtain

$$F(z) \triangleq \sum_{j=0}^{\infty} z^j \Pr[f=j] = \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} z^j \Pr[f=j \,|\, e^*=i] \Pr[e^*=i] = \frac{1}{E'(1)} \sum_{i=1}^{\infty} e(i) \sum_{j=0}^{i-1} z^j \ ,$$

which finally yields

$$F(z) = \frac{E(z)-1}{E'(1)(z-1)} \quad , \tag{II.37}$$

a result that was also reported in [136]. We would like to emphasise that this result remains valid in case of a heterogeneous arrival process, such as the one given by the pgf (II.3), as long as the random ordering assumption, as described above, still holds. An alternative way of reading packets into an output buffer – and one that is encountered in many practical implementations – occurs when packets that arrive during the same slot are forwarded from the inlets into the output buffer in a fixed order, i.e., an arriving packet (if any) that is delivered by inlet i is read into the queue before an arriving packet (if any) that stems from inlet j if $i<j$, $1 \leq i,j \leq N$. In case of a homogeneous arrival process, an arbitrary arriving packet then has an equal probability $1/N$ of being generated by either of the N inlets, which amounts to the random ordering process described above, and implies that the previous expression for $F(z)$ remains valid even under these circumstances. However, when the arrival process is a heterogeneous one that is combined with a fixed ordering of packets that arrive during the same slot (determined by the inlet they originate from), an altered and refined version of the above analysis is mandatory. This issue, among others, will be addressed in Section III.5.3 of the next chapter, and for the time being, we focus on the case where $F(z)$ is given by (II.37). Employment of (II.17a,b) in combination with the relations (A.9) reveals that the first two moments of f are then given by

$$\begin{aligned} \mu_f &= \frac{\sigma_e^2}{2\mu_e} + \frac{\mu_e - 1}{2} \\ \sigma_f^2 &= \frac{\mu_{3,e}}{3\mu_e} - \left(\frac{\sigma_e^2}{2\mu_e}\right)^2 + \frac{\sigma_e^2}{2} + \frac{\mu_e^2 - 1}{12} \end{aligned} \quad . \tag{II.38}$$

In view of the previous remarks and observations, the waiting time w of an arbitrary tagged packet can be written as a function of the queue content $q \equiv q^*$ at the beginning of the arrival slot of this packet and the random variable f associated to this packet, as follows :

$$w = \left\lfloor \frac{f+q}{c} \right\rfloor \quad , \tag{II.39}$$

where, as before, the floor operator $\lfloor x/y \rfloor$ returns the integer part of the fraction x/y. The quantity $f+q$ in this equation equals the number of packets – viewed at the end of the arrival slot – that must be transmitted before the tagged one can be sent, which corresponds to the FCFS queuing discipline under consideration. The division by c, of course, reflects the fact that c packets can be transmitted per slot. Temporarily defining $r \triangleq f+q$, we then deduce from (II.39) that

$$\Pr[w=i]=\sum_{j=0}^{c-1}\Pr[r=ic+j] \quad , \quad i\geq 0 \quad .$$

The pgf $W(z)$ associated with the drv w can therefore be written as

$$W(z)=\sum_{i=0}^{\infty} z^i \sum_{j=0}^{c-1}\Pr[r=ic+j] \quad ,$$

which yields

$$W\left(z^c\right)=\sum_{j=0}^{c-1} z^{-j}\sum_{i=0}^{\infty} z^{ic+j}\Pr[r=ic+j] \quad .$$

Adopting the standard shorthand notation for the discrete Kronecker-delta function

$$\delta(n)=\begin{cases}1 \ , \ n=0\\ 0 \ , \ n\neq 0\end{cases} \quad ,$$

we obtain

$$W\left(z^c\right)=\sum_{j=0}^{c-1} z^{-j}\sum_{i=0}^{\infty}\sum_{k=0}^{\infty} z^k \Pr[r=k]\delta(k-ic-j) \quad .$$

Note that, since $k-j\geq -(c-1)$ in the above expression, the lower bound in the sum for i can be extended to $-\infty$ without affecting the result. In a similar way as in [135] and with $\eta \triangleq \exp\{2\pi\iota/c\}$, we can apply the identity

$$\frac{1}{c}\sum_{n=0}^{c-1}\eta^{mn}=\sum_{i=-\infty}^{\infty}\delta(m-ci) \quad , \tag{II.40}$$

(where ι is the imaginary unit in the complex plane) to eliminate the Kronecker-delta functions in the preceding expression. This equation, in fact, expresses that the sum in the left-hand side is zero, unless m is a multiple of c, in which case it is equal to 1. This results in

$$W\left(z^c\right)=\frac{1}{c}\sum_{j=0}^{c-1}\sum_{n=0}^{c-1}\sum_{k=0}^{\infty}\eta^{n(k-j)}z^{k-j}\Pr[r=k] \quad .$$

Representing by $R(z)$ the pgf of the drv r, this can then be expressed as

$$W\left(z^c\right)=\frac{1}{c}\sum_{j=0}^{c-1}\sum_{n=0}^{c-1}\left(z\eta^n\right)^{-j}R\left(\eta^n z\right)=\frac{1}{c}\sum_{n=0}^{c-1}\frac{1-z^{-c}}{1-\left(\eta^n z\right)^{-1}}R\left(z\eta^n\right) \quad . \tag{II.41}$$

Hence, in view of $R(z)=F(z)Q(z)$, we may finally write

$$W\left(z^c\right) = z^{-c}\sum_{n=0}^{c-1} U_c\left(z\eta^n\right)F\left(z\eta^n\right)Q\left(z\eta^n\right) \quad , \quad U_c(z) \triangleq z\frac{z^c-1}{c(z-1)} \quad , \tag{II.42a}$$

and $U_c(z)$ can be interpreted as being the pgf of a drv, say u_c that is uniformly distributed between 1 and c. In view of expressions (II.42a), (II.37) and (II.13) for $U_c(z)$, $F(z)$ and $Q(z)$ respectively, we finally obtain

$$\boxed{W\left(z^c\right) = z^{-c}\frac{1-\rho}{\rho}\sum_{n=0}^{c-1}\frac{z\eta^n\left(z^c-1\right)}{z\eta^n-1}\frac{E\left(z\eta^n\right)-1}{z^c-E\left(z\eta^n\right)}\Psi\left(z\eta^n\right)} \quad . \tag{II.42b}$$

In addition, appealing to (II.36), we can write

$$\boxed{D\left(z^c\right) = \sum_{n=0}^{c-1} U_c\left(z\eta^n\right)F\left(z\eta^n\right)Q\left(z\eta^n\right)} \quad . \tag{II.43}$$

If we rely on (II.40) to develop the sums in the right-hand sides of the latter two expressions, this leads to the following relations in the probability domain :

$$\begin{cases} w(n) \triangleq \Pr[w=n] = c\Pr[q+f+u_c = c(n+1)] \ , \ n \geq 0 \\ d(n) \triangleq \Pr[d=n] = c\Pr[q+f+u_c = cn] \quad , \ n \geq 1 \end{cases} , \tag{II.44}$$

where $q+f$ is the number of packets that an arbitrary newly arriving packet encounters in the queue, and where u_c is a drv that is uniformly distributed between 1 and c, with mean (c+1)/2, and variance equal to $(c^2-1)/12$. Due to the product form that appears in the right-hand side of (II.42a) and (II.43), u_c can be considered to be statistically independent of q and f; as already pointed out above, q and f are mutually independent drvs as well. These observations provide an insightful probabilistic interpretation to the previous results.

II.4.2 the moments of the packet sojourn time

We will commence by deriving and expression for the mean waiting time μ_w, and the calculations that eventually lead to the final result are presented here in some detail, in order to highlight the methodology that is followed. First, if we apply the operator $\mathcal{M}[\cdot]$ on expression (II.42a), we find

$$\begin{aligned} c\mu_w &= -c + \mu_{u_c} + \mu_f + \mu_q + \mathcal{M}\left[\sum_{n=1}^{c-1} U_c\left(\eta^n z\right)F\left(\eta^n z\right)Q\left(\eta^n z\right)\right] \\ &= -c + \mu_{u_c} + \mu_f + \mu_q + \sum_{n=1}^{c-1}\eta^n U_c'\left(\eta^n\right)F\left(\eta^n\right)Q\left(\eta^n\right) \end{aligned} ,$$

where we have used the property that $U_c(\eta^n)=0$ for $1 \le n \le c-1$. Due to $\eta^{nc}=1$ and expressions (II.42a), (II.37) and (II.13) for $U_c(z)$, $F(z)$ and $Q(z)$ respectively, it is not difficult to show that

$$U_c'\left(\eta^n\right) = \frac{1}{\eta^n - 1} \quad , \quad 1 \le n \le c-1$$

$$F\left(\eta^n\right)Q\left(\eta^n\right) = -\frac{1-\rho}{\rho}\Psi\left(\eta^n\right) = -\frac{1}{c\rho}\sum_{j=0}^{c-1}\eta^{nj}s_c(j) \quad .$$

Consequently, the sum in the right-hand side of the previous expression for $c\mu_w$ can be transformed into

$$\begin{aligned}\sum_{n=1}^{c-1}\eta^n U_c'\left(\eta^n\right)F\left(\eta^n\right)Q\left(\eta^n\right) &= -\frac{1}{c\rho}\sum_{j=0}^{c-1}\sum_{n=1}^{c-1}\frac{\eta^{n(j+1)}}{\eta^n-1}s_c(j) \\ &= -\frac{1}{c\rho}\sum_{j=0}^{c-1}\sum_{n=1}^{c-1}\left\{\frac{1}{\eta^n-1}+\sum_{k=0}^{j}\eta^{nk}\right\}s_c(j) \quad .\end{aligned}$$

On the one hand, we can derive that

$$\sum_{n=1}^{c-1}\frac{1}{\eta^n-1} = \sum_{n=1}^{c-1}\frac{\cos(2\pi n/c)-1-\imath\sin(2\pi n/c)}{\left(\cos(2\pi n/c)-1\right)^2+\left(\sin(2\pi n/c)\right)^2} = \sum_{n=1}^{c-1}\frac{\cos(2\pi n/c)-1}{2\left(1-\cos(2\pi n/c)\right)} = -\frac{c-1}{2} \quad ,$$

while on the other hand, in view of identity (II.40), we find

$$\sum_{n=1}^{c-1}\eta^{kn} = \begin{cases} c-1 \ , \ k=0 \\ -1 \ \ , \ 1 \le k \le c-1 \end{cases} \quad .$$

Hence, making creative use of expression (II.23) for $\Psi(z)$, we deduce that

$$\begin{aligned}\sum_{n=1}^{c-1}\eta^n U_c'\left(\eta^n\right)F\left(\eta^n\right)Q\left(\eta^n\right) &= \frac{1}{c\rho}\sum_{j=0}^{c-1}\left(j-\frac{c-1}{2}\right)\Pr[s \le j] = \frac{1-\rho}{\rho}\left(\Psi'(1)-\frac{c-1}{2}\right) \\ &= \frac{c-\mu_e}{2\mu_e}\sum_{i=1}^{c-1}\frac{1+z_i}{1-z_i} \quad .\end{aligned}$$

If we insert this result in the above equation for $c\mu_w$, and take into account the formulas (II.19) and (II.38) for μ_q and μ_f respectively, then with $\mu_{u_c} = (c+1)/2$, we finally obtain

$$\mu_w = \frac{\sigma_e^2}{2\mu_e(c-\mu_e)} - \frac{1}{2} + \frac{1}{2\mu_e}\sum_{i=1}^{c-1}\frac{1+z_i}{1-z_i} \quad . \qquad \text{(II.45)}$$

Similarly, if we let the operator $\mathcal{V}[\cdot]$ act upon expression (II.42a), then aided by (II.15a), we are able to deduce that

$$c^2\sigma_w^2 = \sigma_{u_c}^2 + \sigma_f^2 + \sigma_q^2 - 2\left(\mu_{u_c} + \mu_f + \mu_q\right)\mathcal{M}\left[\sum_{n=1}^{c-1} U_c\left(\eta^n z\right)R\left(\eta^n z\right)\right]$$
$$+\sum_{n=1}^{c-1}\left(R\left(\eta^n z\right)\left\{\mathcal{V}\left[U_c\left(\eta^n z\right)\right] + \mathcal{M}\left[U_c\left(\eta^n z\right)\right]^2\right\} + 2\mathcal{M}\left[U_c\left(\eta^n z\right)\right]\cdot\mathcal{M}\left[R\left(\eta^n z\right)\right]\right)$$
$$-\left(\sum_{n=1}^{c-1}\mathcal{M}\left[U_c\left(\eta^n z\right)R\left(\eta^n z\right)\right]\right)^2,$$

with $R(z)=F(z)Q(z)$, as before. Adopting similar techniques as above to evaluate the subsequent terms in the right-hand side of this expression eventually yields, after some tedious calculations

$$\sigma_w^2 = \frac{\mu_{3,e}}{3c\mu_e(c-\mu_e)} + \left(\frac{2\mu_e}{c} - 1\right)\left(\frac{\sigma_e^2}{2\mu_e(c-\mu_e)}\right)^2 + \frac{2c\mu_e - 1}{12c^2} - \frac{1}{c^2}\sum_{i=1}^{c-1}\frac{z_i}{(1-z_i)^2}$$
$$-\frac{\sigma_e^2}{2c\mu_e^2}\sum_{i=1}^{c-1}\frac{1+z_i}{1-z_i} - \frac{c^2-\mu_e^2}{4c^2\mu_e^2}\left(\sum_{i=1}^{c-1}\frac{1+z_i}{1-z_i}\right)^2 \qquad \text{(II.46)}$$
$$+2\frac{c-\mu_e}{c^2\mu_e}\sum_{n=1}^{c-1}\left[\frac{1}{\eta^n-1} + \frac{c}{1-E\left(\eta^n\right)} - \sum_{i=1}^{c-1}\frac{\eta^n}{\eta^n - z_i}\right]\frac{\eta^n\Psi\left(\eta^n\right)}{\eta^n - 1}.$$

Equations (II.45) and (II.46) provide us with exact, explicit, and closed-form expressions for the mean and variance of the packet waiting time. Although the result for σ_w^2 looks a bit messy, they lend themselves perfectly well for an implementation in a numerical procedure, once the z_i's, $1\le i\le c$-1, that appear in the expression for $\Psi(\cdot)$ (and its coefficient), have been calculated numerically by some root-finding algorithm, such as a Newton-Raphson scheme. One can easily verify that, notwithstanding the presence of the complex numbers η^n in these expressions, the final results for μ_w and σ_w^2 are real-valued quantities, as should evidently be the case. Also note that in the course of these derivations, we have tacitly assumed that $E(\eta^n)\neq 1$ for $1\le n\le c-1$. This will be so if $\mathcal{L}=1$, with $\mathcal{L}$ the hcf of the set of integers $\{\{c\}\cup\{n\in\mathbb{N}:\Pr[e=n]\neq 0\}\}$. Then the property $z_i\neq\eta^n$, $1\le i\le c-1$, is also valid, thus prohibiting that some of the terms in the above expression become infinite. If $E(\eta^n)\neq 1$ does not hold for some values of n, then the foregoing derivations can be adjusted without significant difficulty to capture this phenomenon.

The mean and variance of the packet delay d immediately follow from (II.36), i.e.,

$$\begin{aligned} \mu_d &= \mu_w + 1 \\ \sigma_d^2 &= \sigma_w^2 \end{aligned} \quad . \tag{II.47}$$

In view of these results and the ones established in Section II.3.2.1, the mean packet waiting time and delay apparently satisfy

$$\begin{aligned} \mu_e \cdot \mu_w &= \mu_q \\ \mu_e \cdot \mu_d &= \mu_s \end{aligned} \quad ,$$

which confirms that these results concur with Little's theorem (e.g. (I.5)), applied on the queue and the entire buffer system respectively, as they should.

If we want to avoid the calculation of the zeroes z_i ($1 \le i \le c$-1) altogether, then we can combine the results of section II.3.2.2 (see equation (II.25b)) to establish that

$$0 \le \frac{1}{2c\rho} \sum_{i=1}^{c-1} \frac{1+z_i}{1-z_i} \le \frac{(c-1)}{2c(1-\rho)} + m\frac{(m+1-2c\rho)}{2c^2(1-\rho)\rho} \quad , \tag{II.48}$$

and inserting this result into (II.45) (and (II.47)) leads to a lower and upper bound for μ_w (and μ_d). Note that the upper bound in (II.48) does not tend to infinity as ρ (and μ_e)$\rightarrow$0, since $m = \lfloor c\rho \rfloor$ is equal to zero as well for sufficiently small values of ρ. In addition, approximation (II.28) for $\mathcal{M}[\Psi]$can be invoked to yield an accurate approximation for the sum in the above expression that contains the z_i's. Finding an appropriate approximation and upper- and lower bounds for σ_w^2 (and σ_d^2) is a totally different matter and far from straightforward, in view of the rather complicated combination of terms that contain the z_i's in the right-hand side of (II.46). To address this issue, we refer to Section III.4.3, where we will present a closer examination of the packet waiting time and delay pmfs that yet allows us produce acceptable bounds for the variance of these quantities.

We conclude this section by considering the heavy-load limits of the mean and variance of the packet waiting time and delay, which now satisfy

$$\begin{aligned} \mu_{w,\rho\rightarrow 1} &= \mu_{d,\rho\rightarrow 1} = \frac{\sigma_e^2}{2\mu_e(c-\mu_e)} \\ \sigma_{w,\rho\rightarrow 1}^2 &= \sigma_{d,\rho\rightarrow 1}^2 = \left(\frac{\sigma_e^2}{2\mu_e(c-\mu_e)}\right)^2 . \end{aligned} \tag{II.49}$$

Hence, analogous to the results of Section II.3.2.3, the heavy-load behaviour of both the mean and variance of the packet waiting time and delay is determined by the variance of the packet arrival process as well. In addition, we also observe that the coefficient of variation of these drvs, which equals the ratio of their stdv versus their mean, tends to 1 as $\rho\rightarrow$1.

II.4.3 tail distribution of the packet sojourn time

Let us first derive an expression for the asymptotic behaviour of the packet waiting time, where we depart from the pgf $W(z^c)$ given by (II.42b). These deductions build upon the results of Appendix A, and we follow a similar approach as in [6]. A crucial initial observation is that this expression for $W(z^c)$ does not satisfy the condition that it has only one pole for which the modulus is minimal. Indeed, given that z_o is the solution of $z^c - E(z)=0$ in the interval $]1, \mathcal{R}[$, as before, then each of the denominators $z^c - E(z\eta^n)$ in the expression for $W(z^c)$ have a root with modulus z_o, namely the complex number $z_o\eta^{-n}$, for all $0 \le n \le c-1$, and these c (dominant) poles will determine the asymptotic behaviour of the drv w. This implies that the method that was outlined in Appendix A must be extended to deal with this situation, by considering a term similar to (A.13a,b) for each of these poles, leading to

$$\Pr[w=i] \cong \sum_{n=0}^{c-1} \left\{ \lim_{z \to z_o \eta^{-n}} \left(1 - \frac{z}{z_o \eta^{-n}} \right) W\left(z^c\right) \right\} \left(z_o \eta^{-n}\right)^{-ic} \quad ,$$

where we have used the property that $\Pr[w=i]$ is the coefficient of z^{ic} in the expression for $W(z^c)$. From (II.42b), it can be shown that

$$\lim_{z \to z_o \eta^{-n}} \left(1 - \frac{z}{z_o \eta^{-n}} \right) W\left(z^c\right) = -\frac{1-\rho}{\rho} \frac{z_o^{-c}\left(z_o^c - 1\right)}{z_o - 1} \frac{E(z_o)-1}{c z_o^{c-1} - E'(z_o)} \Psi(z_o) \triangleq \frac{\varsigma}{c} , \qquad \text{(II.50a)}$$

and with this definition of ς, we obtain

$$\Pr[w=i] \cong \frac{\varsigma}{c} \sum_{n=0}^{c-1} \eta^{nic} z_o^{-ic} = \varsigma z_o^{-ic} \quad ,$$

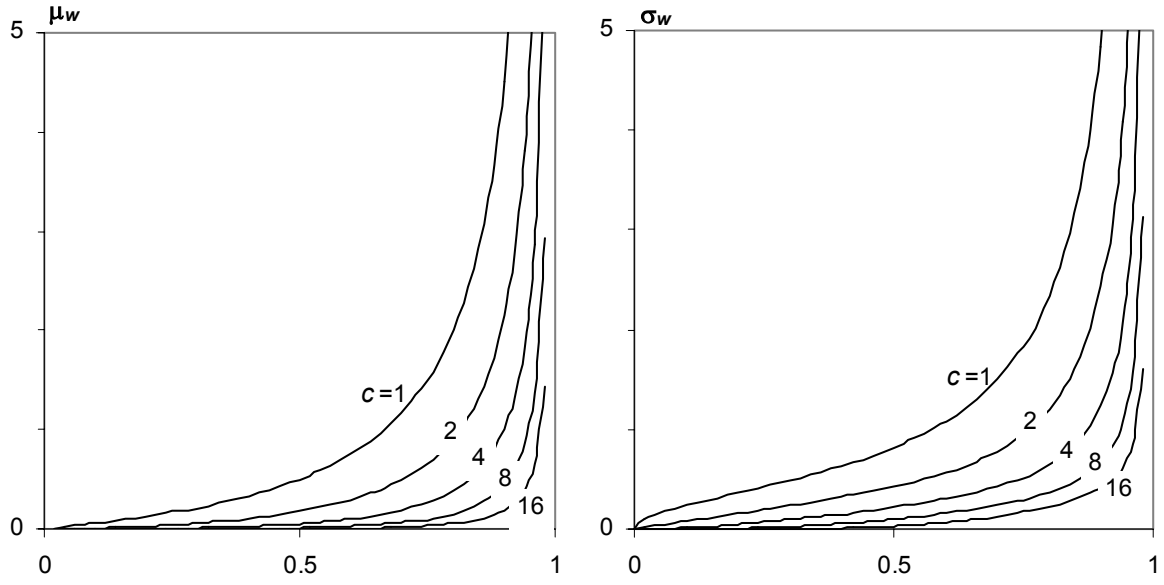
where we have once again invoked the identity (II.40). The probability that the drv w exceeds a given threshold T_h then becomes

$$\Pr[w > T_h] \cong \frac{\varsigma}{z_o^c - 1} z_o^{-T_h c} \quad , \qquad \text{(II.50b)}$$

for sufficiently large values of T_h. Since $\Pr[d > T_h] = \Pr[w > T_h - 1]$, for all $T_h \ge 1$, we also find

$$\Pr[d > T_h] \cong -\frac{\varsigma}{z_o^c - 1} z_o^{-(T_h - 1)c} \quad . \qquad \text{(II.50c)}$$

These results indicate that, if the asymptotic tail behaviour of the queue and system content is determined by their dominant pole z_o, then z_o^c will determine the tail behaviour of the pmf and ccdf of the drvs w and d that characterise an arbitrary packet's sojourn time in the buffer.

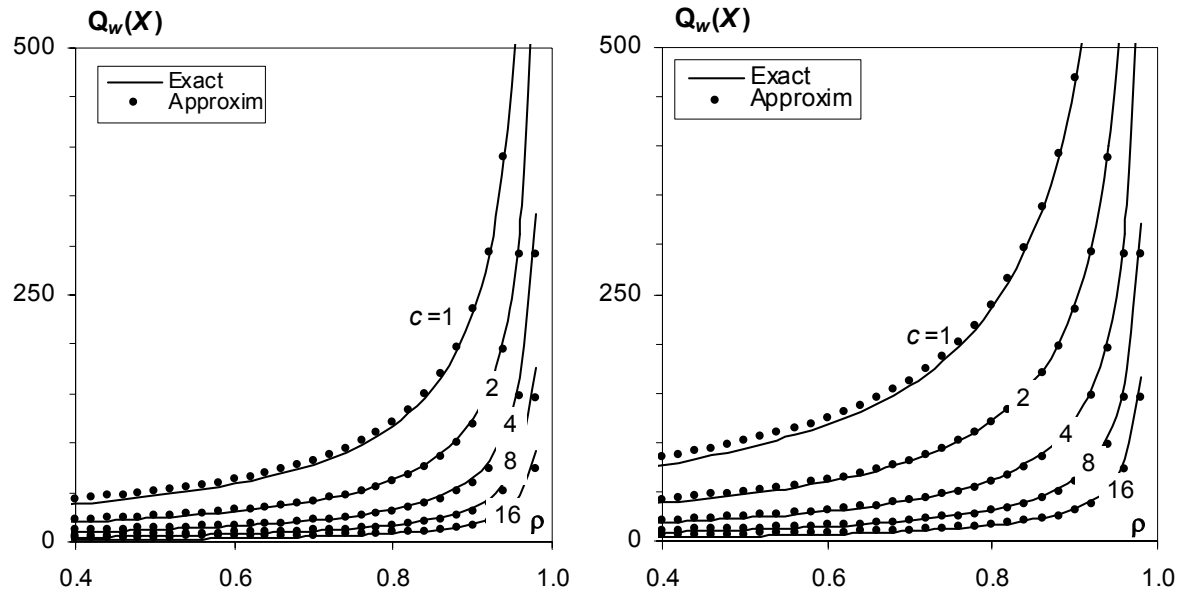


(a) c=1 (b) c=16

Figure II-14 : moments of the waiting time versus ρ; Poisson arrival process

II.4.4 numerical examples

The performance indices concerning the buffer content discussed in Section II.3, such as the mean, stdv and 10^{-X} quantile, revealed that these only weakly depend on the number of output links per buffer, represented by the value of c. This is no longer true as far as the packet sojourn time is concerned, which strongly depends on the number of packets that can be transmitted per slot. This is illustrated in Fig. II-14, where we have plotted the mean and stdv of the packet waiting time versus ρ, for a Poisson packet arrival process, and c=1,2,4,8,16. Apparently, these quantities strongly decrease for increasing values of c. This for instance also follows from the heavy-load limits of the mean and variance of the waiting time in expression (II.49), which can serve as a back-of-the-envelope approximation for μ_w and σ_w^2 respectively. Starting from c=1 and any i.i.d. arrival pattern generated by N sources, then if we increase c and increase the number of (independent) sources that generate packet arrivals accordingly, the numerator σ_e^2 in the right-hand side of these expressions will be proportional to c (and/or N), while the corresponding denominators clearly are quadric functions of c. Hence, the steep decline of the value of these moments observed in Fig. II-14 as c becomes larger. This comprises the main advantage of employing multi-server queuing systems : whereas the gain with respect to the amount of customers in the system is very limited – or even nonexistent – by increasing the number of servers, the delays that these customers will experience are significantly reduced, for comparable arrival patterns.



(a) X=4 (b) X=8

Figure II-15 : quantile of the waiting time versus ρ; clumped arrival process with κ_c=5

These observations are supported by the behaviour of the 10^{-X} waiting time quantiles, that are depicted in Figs. 15a,b versus ρ for X=4 and 8 respectively and a clumped packet arrival process with κ_c=5. Similar to (II.35a), expression (II.50a) provides an efficacious, albeit approximate, way for calculating these quantities, leading to

$$Q_w(X) = \ln\left(z_o^c\right)^{-1}\left(\ln(\varsigma) - \ln\left(z_o^c - 1\right) + X\ln(10)\right) \quad . \tag{II.50a}$$

For increasing c, i.e., c=1,2,4,8,16, we can only conclude from Figs II-15a,b that the $Q_w(X)$ performance characteristic strongly decreases as well, which is exemplified by the dependence on z_o^c of the above expression. These figures also provide a comparison between the 'exact' quantile values (II.50a), and an approximation that relates these quantities with the first two (central) moments of the drv w, which, in a similar manner as in (II.35b), is calculated as

$$Q_w(X) \cong 1 + \mu_w + (X.\ln(10) - 1)\sigma_w \quad . \tag{II.50b}$$

Once again, we perceive a remarkable resemblance between the two approaches, as long as the load ρ is not too low.

II.5 Buffers with finite size

When we are dealing with buffers of finite size, the loss process of the packets that are rejected due to buffer overflow becomes an important issue. In this section, we examine the loss process of a finite-capacity buffer by means of a generating-functions approach, and try to refine and/or extend existing results. The purpose is to express some of the crucial performance measures that are related to a finite-capacity buffer, in terms of the performance measures associated with its infinite-capacity counterpart. This is a challenging issue that hopefully leads to new efficient calculation methods for the former quantities, since the latter quantities are often available under some (semi-)analytic form, as demonstrated in the preceding sections. In the existing literature, quantities such as the packet loss ratio are often approximated by some tail probability of the buffer content in the corresponding infinite-capacity queuing model ([43], [124], [153], [170], [185], [192], [194], [217], [243], [262], [271], [274]); however, how these two quantities are related, is generally not very well known, in particular when multi-server buffers are involved. In [193], a fluid model, that neglects the specific packetised nature of the information flow, is investigated for the purpose of establishing such a relationship. This topic, for a packet-based arrival process, is exactly the focus of this section, and builds on the work presented in [43].

II.5.1 loss due to buffer overflow

Let us investigate this issue in more detail, and, as before, consider a buffer system with c ($c \geq 1$) output links, implying that the number of packets that can leave the system during any slot is at most c. As depicted in Fig. II-16, the queue, where packets that await their transmission are temporarily stored, is assumed to have a storage capacity of K_q packets; this does not include the packets that are being transmitted during a slot, if any. As discussed in Section I.2.2, newly arriving packets do not get immediate access to the transmission units of the system (i.e., the c servers) during their arrival slot – even when some of these are idle – but are initially stored in the queue; hence, they may enter the transmission part of the buffer system at the earliest during the slot following their arrival slot. We adopt a *drop-tail* (DT) buffer management policy, implying that new packet arrivals are accepted as long as buffer space is available in the queue upon arrival; otherwise, packets will be dropped.

Analogous to the foregoing analyses, we denote by $\tilde{s}_n$ (with pgf $\tilde{S}_n(z)$) the system content (the number of packets in the system, including those that are currently being transmitted) at the beginning of slot n, and yet again we let the drv e_n (with pmf $e(\cdot)$, cdf $e_c(\cdot)$, and pgf $E(z)$) characterise the number of new packet arrivals that attempt to enter the finite-capacity buffer during slot n, which is an i.i.d. process by assumption. Since we observe the system content just after possible departure epochs, but before new arrivals, then in view of the previous, this quantity equals the amount of packets in the queue at the end of the preceding slot, and

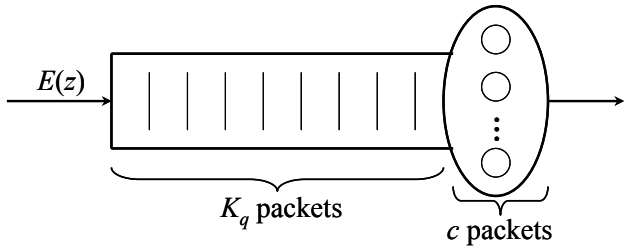


Figure II-16 : storage capacity in a finite buffer system

therefore it cannot exceed the queue capacity K_q. Consequently, the system content at (i.e., just after) consecutive slot boundaries is governed by the system equation (e.g. (I.2a))

$$\tilde{s}_{n+1} = \min\left\{(\tilde{s}_n - c)^+ + e_n, K_q\right\} \quad . \tag{II.51}$$

Calculating quantities such as the pmf of the system content involves solving the K_q+1 balance questions (together with the normalisation condition) that derive from (II.51). Indeed, defining the steady-state pmf and cdf of the buffer content

$$\tilde{s}(j) \triangleq \lim_{n\to\infty} \Pr[\tilde{s}_n = j] \; ; \; \tilde{s}_c(j) \triangleq \lim_{n\to\infty} \Pr[\tilde{s}_n \le j] \; , \; 0 \le j \le K_q \quad ,$$

then the cdf $\tilde{s}_c(j)$ (and hence, the pmf $\tilde{s}(j)$) can be calculated numerically from the set of linear equations

$$\tilde{s}_c(j) - \sum_{i=0}^{(j,K_q-c)^-} \tilde{s}_c(i+c)e(j-i) = 0 \; , \; 0 \le j \le K_q - 1 \tag{II.52}$$

(where $(x,y)^- \triangleq \min\{x,y\}$), together with the identity $\tilde{s}_c(K_q) = 1$, that follows from the normalisation of the pmf.

The calculation of this pmf allows us to study the loss process as well. Let us denote by ℓ_n (with pgf $L_n(z)$) the number of packets lost during slot n, due to contention when newly arriving packets arrive at a queue that is already fully occupied. This drv is related to $\tilde{s}_n$ and e_n by the system equation (e.g. (I.2b))

$$\ell_n = \left((\tilde{s}_n - c)^+ + e_n - K_q\right)^+ \quad , \tag{II.53}$$

which expresses that, if the number of old and newly arriving packets in the queue exceeds K_q, then the excess packets will all be rejected; otherwise, they are able to enter the system. Under steady-state conditions, we let ℓ (with pmf $\ell(i)$) represent the number of packets that are lost during an arbitrary slot. Then the statistical independence of $\tilde{s}_n$ and e_n, and (II.53), imply that

$$\ell(i) = \sum_{j=0}^{c-1} \tilde{s}(j) e\left(K_q + i\right) + \sum_{j=c}^{K_q} \tilde{s}(j) e\left(K_q + c + i - j\right) \quad , \quad i \geq 1 \quad . \tag{II.54}$$

From this formula, quantities such as the mean μ_ℓ and variance of the number of packets lost per slot can be readily derived. We should also point out that, since we assume a steady-state environment, then μ_ℓ can also be calculated as the difference between the mean number of packet arrivals, (i.e., $c\rho$) and the average number of packets that leave the buffer during an arbitrary slot

$$\begin{aligned} \mu_\ell &= c\rho - \left[\sum_{j=0}^{c-1} j\tilde{s}\left(j\right) + c \sum_{j=c}^{K_q} \tilde{s}\left(j\right) \right] = \sum_{j=0}^{c-1} \left(c - j\right)\tilde{s}\left(j\right) - c\left(1-\rho\right) \\ &= \sum_{j=0}^{c-1} \tilde{s}_c\left(j\right) - c\left(1-\rho\right) \ . \end{aligned} \tag{II.55}$$

Solving the set of balance equations and calculating quantities such as μ_ℓ may become a numerically demanding task, depending on the value of K_q and the values that the $\tilde{s}\left(j\right)$'s assume : typically, whenever K_q is high, and/or the values of the $\tilde{s}\left(j\right)$'s differ many orders of magnitude, substantial numerical problems may arise. On the other hand, the aforementioned references indicate that, under a wide range of circumstances, the packet loss ratio (P_{LR}) – which is defined as the fraction of newly arriving packets that are lost due to buffer overflow and therefore equals $\mu_\ell / (c\rho)$ – in a buffer with finite storage capacity, and the ccdf of the buffer content in the corresponding infinite-capacity queue, exhibit a similar asymptotic behaviour when plotted versus K_q. Therefore, for a lot of cases, P_{LR} is approximated by the latter quantity, since the results of Section II.3.3 demonstrate that the asymptotic behaviour of the ccdf of the buffer content is fairly easy to capture and can be calculated in an efficient way by the dominant pole approximation. Nevertheless, the exact nature of correspondence between the number of packets lost per slot and the distribution of the buffer content in a buffer with infinite storage capacity remains obscure, and this issue has received little or no attention in the literature. In [128], a relation between the P_{LR} in the discrete-time $Geo^{[X]}/D/1/K_q$ queuing system, and the distribution of the buffer content in the corresponding infinite capacity queuing model was reported, which is a special case of the multi-server system studied in these sections.

II.5.2 probability generating function of the buffer content

The system equation (II.53) that characterises the loss process can be z-transformed into

$$L_n(z) = 1 + \mathbf{E}\left[\left(z^{(\tilde{s}_n - c)^+ + e_n - K_q} - 1\right)\left\{(\tilde{s}_n - c)^+ + e_n > K_q\right\}\right] \quad .$$

On the other hand, from system equation (II.51) we deduce that

$$\tilde{S}_{n+1}(z) = \mathbf{E}\left[z^{(\tilde{s}_n - c)^+ + e_n}\right] + \mathbf{E}\left[z^{K_q} - z^{(\tilde{s}_n - c)^+ + e_n}\left\{(\tilde{s}_n - c)^+ + e_n > K_q\right\}\right] \quad .$$

These two equations can be converted into the relation

$$\tilde{S}_{n+1}(z) = \mathbf{E}\left[z^{(\tilde{s}_n - c)^+ + e_n}\right] + z^{K_q}\left(1 - L_n(z)\right) \quad ,$$

which, from a similar analysis as in Section II.3.1, can be further written as

$$\tilde{S}_{n+1}(z) = z^{-c}E(z)\left(\tilde{S}_n(z) + \sum_{j=0}^{c-1}\left(z^c - z^j\right)\tilde{s}(j)\right) + z^{K_q}\left(1 - L_n(z)\right) \quad .$$

The finite-buffer system reaches a stochastic equilibrium after a sufficiently large period of time, and the pmf (and corresponding pgf) of all random variables involved in this analysis becomes time-independent (i.e., independent of n). Then, defining the steady-state pgfs

$$\tilde{S}(z) \triangleq \lim_{n\to\infty} \tilde{S}_n(z) \ ; \ L(z) \triangleq \lim_{n\to\infty} L_n(z) \quad ,$$

we obtain the following expression for the steady-state pgf of the system content at the beginning of a random slot

$$\tilde{S}(z) = \frac{\left(c(1-\rho) + \mu_\ell\right)(z-1)E(z)\tilde{\Psi}(z) + z^{K_q}\left(1 - L(z)\right)}{z^c - E(z)}$$

$$\tilde{\Psi}(z) \triangleq \frac{1}{c(1-\rho) + \mu_\ell}\sum_{j=0}^{c-1}\frac{z^c - z^j}{z-1}\tilde{s}(j) = \frac{1}{c(1-\rho) + \mu_\ell}\sum_{j=0}^{c-1} z^j \tilde{s}_c(j) \quad . \qquad \text{(II.56)}$$

In view of this definition and the property that $\mu_\ell \equiv L'(1)$ represents the mean number of packets lost per slot, we deduce that $\tilde{\Psi}(1) = 1$, in concurrence with (II.55). The buffer content at the beginning of a slot is bounded by the value of K_q; therefore, the right-hand side in the above formula must be a polynomial in z of degree K_q, implying that the denominator must be a divisor of the numerator. As we will demonstrate in the next section, this property can be further exploited to derive useful formulas for the characteristics of the loss process from this expression.

II.5.3 pgf and moments of the loss process

If we combine expression (II.10) for $S(z)$ with (II.56), we can establish the following expression that relates the finite- and infinite-size buffer characteristics

$$\Psi(z)\tilde{S}(z) = \left(1 + \frac{\mu_\ell}{c(1-\rho)}\right)\tilde{\Psi}(z)S(z) + \frac{1-L(z)}{z^c - E(z)}z^{K_q+c}\Psi(z) \quad .$$

Since we make use of infinite-buffer size results in this expression, we assume that the associated equilibrium condition $\rho<1$ is fulfilled. The left-hand side of this expression is a polynomial in z of degree K_q+c-1, while the power series expansion around $z=0$ of the second term in the right-hand side will merely contain terms whose degree of z is at least K_q+c. Consequently, taking into account the definitions of $\tilde{\Psi}(z)$ and $S(z)$ and by explicitly considering the power series expansion around $z=0$ of this term, we can derive the following formula for $L(z)$, the equilibrium pgf of the number of packets lost per slot :

$$L(z) = 1 + \frac{z-1}{Q(z)}\left\{\sum_{j=1}^{c-1}\sum_{i=0}^{j-1} z^i \tilde{s}_c(j)s\left(K_q + c + i - j\right) + \left(c(1-\rho)+\mu_\ell\right)\tilde{\Psi}(z)\sum_{i=0}^{\infty} z^i s\left(K_q + c + i\right)\right\}, \quad \text{(II.57)}$$

where, as before, $Q(z)$ represents the steady-state pgf of the queue content in a buffer with infinite storage capacity, and where $s(\cdot)$ and $s_c(\cdot)$ represent the pmf and cdf of the corresponding system content.

From this expression for $L(z)$, all quantities of interest concerning the number of packets lost per slot can be calculated. For instance, the loss probability P_{LP}, which we define as the probability that packet loss occurs during a slot, is equal to $1-L(0)$, and therefore given by

$$P_{LP} = \frac{E(0)}{c(1-\rho)\Psi(0)}\sum_{j=0}^{c-1}\tilde{s}_c(j)s\left(K_q + c - j\right) \quad .$$

Taking the first derivative with respect to z for $z=1$ of both hand sides of (II.57) produces the following expression for μ_ℓ :

$$\mu_\ell = \frac{1}{s_c(K+c-1)}\left\{\sum_{j=1}^{c-1}\sum_{i=0}^{j-1}\tilde{s}_c(j)s\left(K_q + c + i - j\right) + c(1-\rho)\left(1 - s_c\left(K_q + c - 1\right)\right)\right\} .$$

In addition, applying the $\mathcal{V}[\cdot]$ operator on the previous expression for $L(z)$ yields a formula for the variance of the number of packets lost per slot :

$$\sigma_\ell^2 = \mu_\ell\left(1 - \mu_\ell - 2\mu_q\right) + 2\left\{\sum_{j=1}^{c-1}\sum_{i=0}^{j-1} i\tilde{s}_c(j)s\left(K_q + c + i - j\right) + \sum_{j=0}^{c-1}\sum_{i=0}^{\infty}(i+j)\tilde{s}_c(j)s\left(K_q + c + i\right)\right\} .$$

These results allow the (theoretical) calculation of P_{LP} and the mean and variance of the number of packets lost per slot, but are not very practical since they contain the probabilities $\tilde{s}_c(j)$, $0 \le j \le c\text{-}1$, and $s(i)$, $i > K_q + c$. It is self-evident to approximate the latter probabilities by their dominant-pole approximation given by (II.32), while we approximate $\tilde{\Psi}(z)$ by $\Psi(z)$ in order to deal with the $\tilde{s}_c(j)$'s. Indeed, the system parameters should be tuned such that packet loss is a rare event, in which case we expect this approximation for $\tilde{\Psi}(z)$ to be sufficiently accurate. If we define

$$\tau(K_q) \triangleq \frac{\Psi(z_o)\dfrac{\zeta z_o^{-K_q+1}}{z_o - 1}}{1 - \Psi(z_o)\dfrac{\zeta z_o^{-K_q+1}}{z_o - 1}} \quad , \tag{II.58a}$$

then this approach yields the following results

$$\begin{aligned} P_{LP} &\cong \frac{E(0)}{\psi(0)}\left(1 - z_o^{-1}\right)\tau(K_q) \\ \mu_\ell &\cong c(1-\rho)\tau(K_q) \\ \sigma_\ell^2 &\cong \mu_\ell\left(1 - \mu_\ell - 2\mu_q\right) + 2c(1-\rho)\frac{\tau(K_q)}{z_o - 1} \; . \end{aligned} \tag{II.58b}$$

Moreover, since packet loss is (supposed to be) a rare event, it is also interesting to take a look at the conditional packet loss process, given that packet loss occurs. If we denote by the drv ℓ_c the number of packets that are lost, conditioned on the event that packet loss occurs in the same slot, then the pgf of this quantity is given by $(L(z)\text{-}1+P_{LP})/P_{LP}$, and its mean and variance satisfy

$$\begin{aligned} \mu_{\ell_c} &= \mu_\ell / P_{LP} \\ \sigma_{\ell_c}^2 &= \sigma_\ell^2 / P_{LP} - \mu_{\ell_c}^2 \left(1 - P_{LP}\right) \; . \end{aligned} \tag{II.59}$$

These expressions are easy to evaluate numerically, especially when combined with the approximations (II.27) and (II.25b) for $\Psi(z)$ and the sum containing the z_i's that appears in the expression for μ_q respectively. If we compare expressions (II.58a,b) for μ_ℓ with the ones for $\Pr[q \ge K_q]$ (i.e., (II.31b,c)), we notice that, apart from the factor $\Psi(z_o)$, these two quantities exhibit the same kind of behaviour if K_q is sufficiently large, in which case the denominator in expression (II.58a) for $\tau(K_q)$ is practically equal to 1. This explains why the packet loss ratio

$\mu_\ell/(c\rho)$ in a buffer of size K_q is often approximated by the tail probability $\Pr[q \geq K_q]$; the above formulas however indicate how their relation can be refined.

Although these results have been established under the condition that the offered load satisfies $\rho < 1$, a remarkable and important feature is that they remain valid for values of $\rho > 1$ as well. Under these circumstances, the dominant pole z_o – which is still equal to the radius of convergence of $Q(z)$ (and $S(z)$) – falls within the complex unit circle, i.e., $|z_o| < 1$, and it is not difficult to check that $\tau(K_q)$ converges to -1 as K_q increases for such values of ρ. For instance, as far as μ_ℓ is concerned, this shows that this quantity evolves to $c\rho$-c for sufficiently high values of K_q : since the system is overloaded, the queue is (practically) never empty under this traffic scenario, implying that c packets can (almost always) be transmitted during every slot, and because on average $c\rho$ packets are offered per slot, the difference $c\rho$-c will indeed become equal to the packet loss intensity μ_ℓ for large enough values of K_q. The behaviour of the above formulae for P_{LP} and σ_ℓ is less clear if $\rho > 1$; however, the accuracy of results for $\rho < 1$ as well as $\rho > 1$ will be amply illustrated in Section II.5.5.

II.5.4 loss and loss-free periods

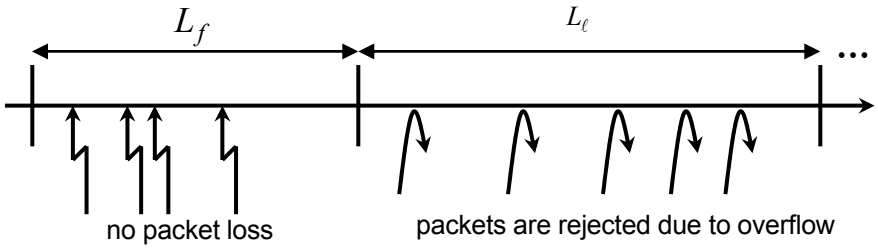
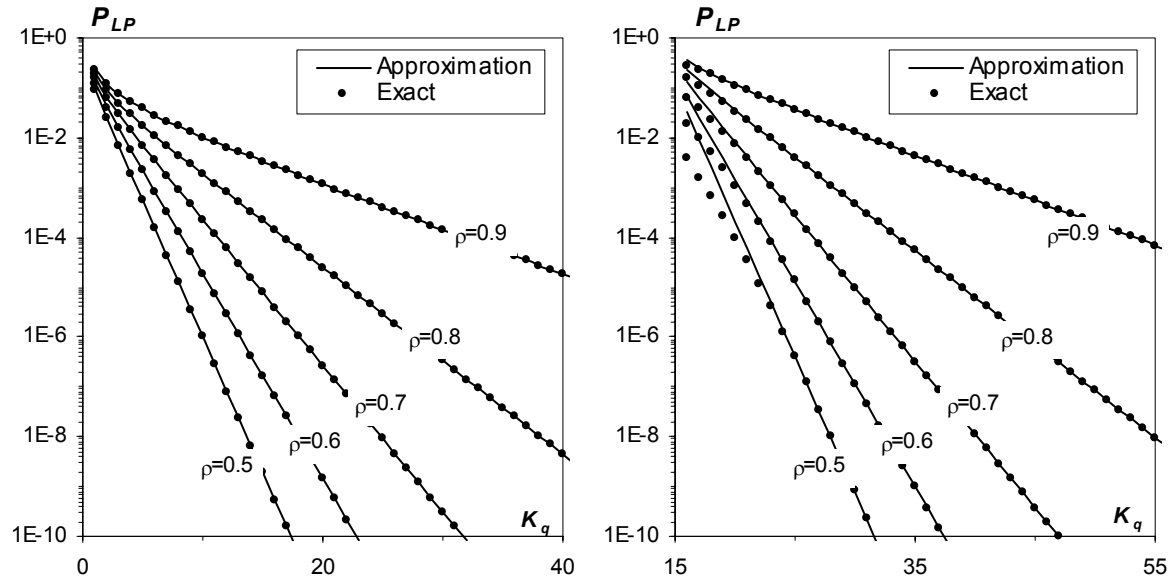


Figure II-17: alternating loss-free and loss periods

In addition to studying the number of packets lost per slot, it is interesting as well to take a quick look at how packet loss is spread in time. For that purpose, the time axis can be divided into alternating periods during which packet loss occurs, and no loss occurs respectively. As illustrated in Fig. II-17, we denote by L_ℓ and L_f the length of these subsequent *loss* and *loss-free periods*, and the purpose of this section is to study these two quantities into somewhat more detail.

Let ℓ once again represent the number of packets lost during a randomly chosen slot, and the drv ℓ_- the number of packets that were lost during the preceding slot. If ℓ_- is nonzero, then the buffer content at the beginning of the tagged slot perforce equals K_q, and the (conditional) pmf of ℓ is therefore given by, for $j>0$,

$$\Pr[\ell = i \,|\, \ell_- = j] = \begin{cases} e(i+c) \ , \ i \geq 1 \\ e_c(c) \ \ , \ i = 0 \end{cases} \quad , \quad j>0 \quad .$$



(a) c=1 (b) c=16

Figure II-18 : loss probability versus K_q; Poisson arrival process

Provided that $\ell\_ > 0$, these probabilities do not depend on the value of $\ell\_$; hence, we obtain

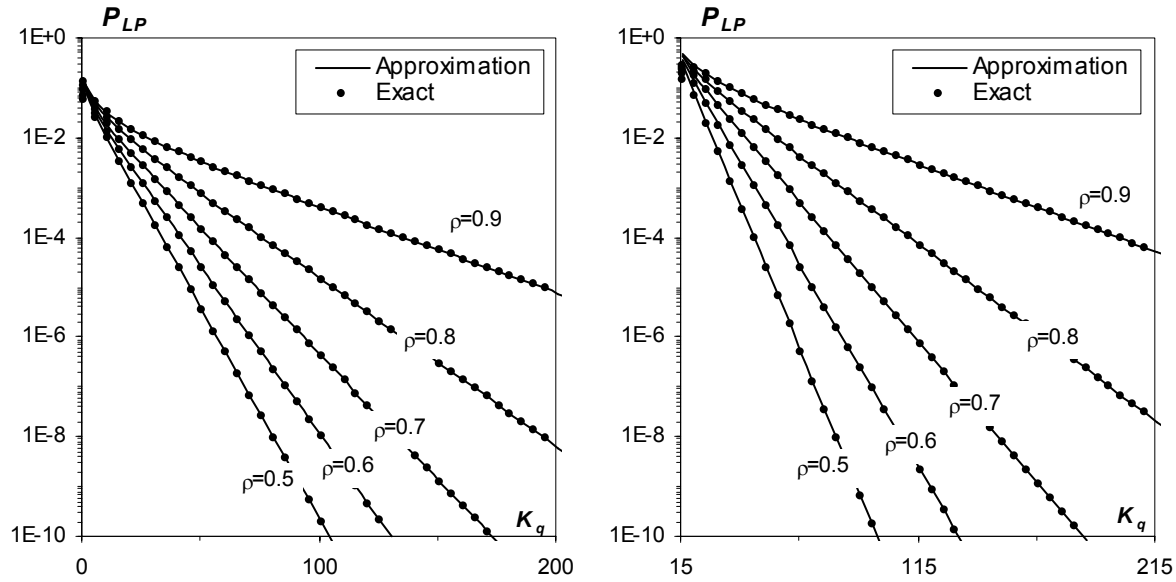
$$\begin{cases} \Pr[\ell > 0 | \ell\_ > 0] = 1 - e_c(c) \\ \Pr[\ell = 0 | \ell\_ > 0] = e_c(c) \end{cases},$$

implying that the drv L_ℓ is geometrically distributed with parameter 1-$e_c(c)$, and mean $1 / e_c(c)$,

$$\Pr[L_\ell = n] = e_c(c)(1 - e_c(c))^{n-1} \ , \ n \geq 1 \ .$$

Due to the memoryless and independent nature of these geometrically distributed loss periods, it becomes clear that the lengths of successive loss and loss-free periods form a series of statistically independent random variables. The characteristics of the length of a loss-free period are somewhat harder to determine, and depend on the details of the packet arrival process during such a period (and the last slot of the loss period prior to such a loss-free period). Nevertheless, an expression for the mean length of a loss-free period can be derived from the observation that the packet loss probability P_{LP}, as defined in the previous section, must be equal to the fraction of slots that belong to a loss period, which implies

$$\mathbf{E}[L_f] = \frac{1 - P_{LP}}{P_{LP}} \mathbf{E}[L_\ell] \ .$$



(a) c=1 (b) c=16

Figure II-19 : loss probability versus K_q; clumped arrival process with κ_c=5

We can conclude by taking a brief look at the number of packets that are lost during the consecutive slots of a loss period. If we denote by $\ell_{\bar{1}}$ the number of packets lost during any slot of a loss period, except for the first slot, then the mean value of this quantity satisfies

$$\mathbf{E}\left[\ell_{\bar{1}}\right] \triangleq \mathbf{E}[\ell|\ell>0,\ell_{-}>0] = \frac{1}{1-e_c(c)}\mathbf{E}[\ell|\ell_{-}>0] = \frac{1}{1-e_c(c)}\sum_{i=c+1}^{\infty}(i-c)e(i)$$

$$= \frac{1}{1-e_c(c)}\left\{\sum_{i=0}^{c-1}e_c(i) - c(1-\rho)\right\} .$$

Consequently, if we define ℓ_1 as the number of packets lost during the first slot of a loss period, then the expected value of this quantity readily follows from the observation that μ_{ℓ_c} represents the mean number of packet losses during a random slot of a loss period; hence

$$\mathbf{E}[\ell_1] \triangleq \mathbf{E}[\ell|\ell>0,\ell_{-}=0] = \frac{\mu_{\ell}}{e_c(c)P_{LP}} + \frac{1-e_c(c)}{e_c(c)}\mathbf{E}\left[\ell_{\bar{1}}\right] \quad .$$

II.5.5 numerical examples

In Figs. II-18a,b, we illustrate the accuracy of the approach that was expounded upon in the foregoing sections, by comparing the exact values of P_{LP}=1−$\ell(0)$ (as a function of K_q, with $K_q \geq c$), calculated by means of (II.54), with the approximation given by (II.58b). The packet arrival process is a Poisson process, and the number of transmission lines per buffer equals 1 (a) and 16 (b) respectively. From these curves, we can only conclude that the proposed

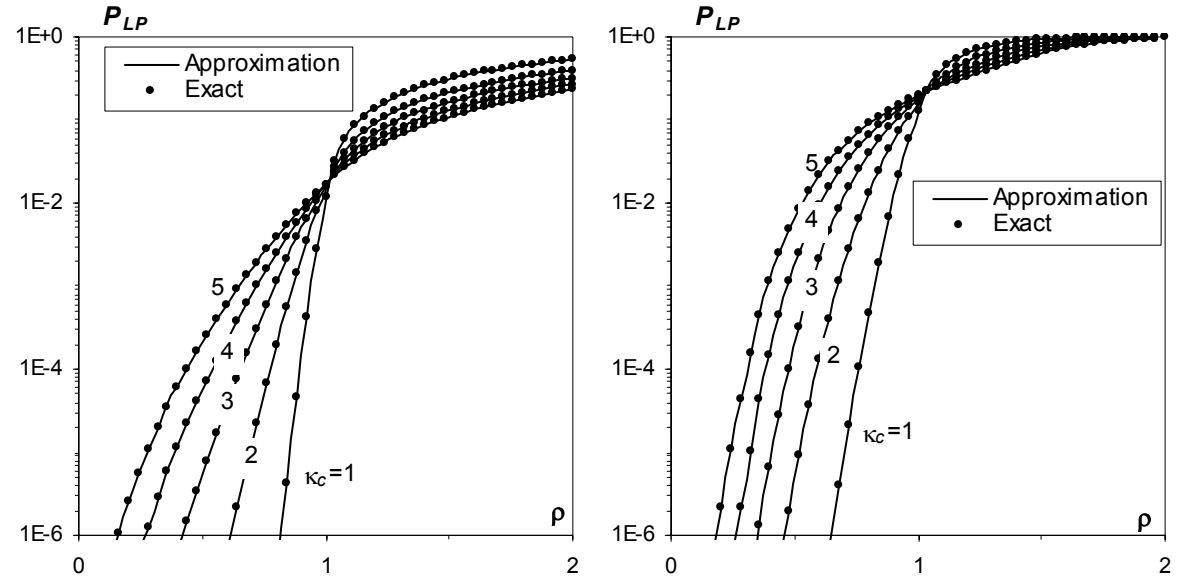
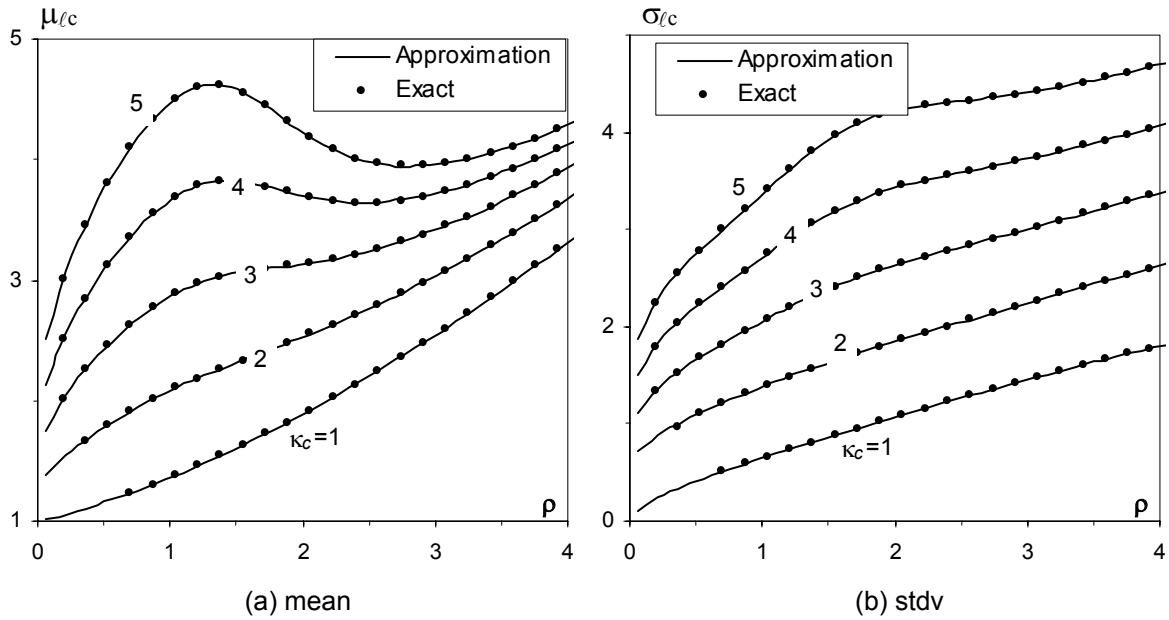


Figure II-20 : loss probability versus ρ; buffer size K_q=30

approximate approach is indeed highly accurate, unless ρ and K_q are low, in which case some discrepancy between the two methodologies can be observed. That these conclusions are not limited to the Poisson-arrivals case can be viewed in Figs. II-19a,b, where we have plotted similar curves, for a clumped packet arrival process with κ_c=5. Since the calculation of the exact results can be very demanding for large values of K_q – and indeed is bound to become unfeasible at a certain point – the advantage of adopting the approximate technique is obvious. Moreover, comparison of Figs. II-18a,b and II-19a,b also points to the fact that, for a given targeted packet loss probability, the minimal buffer space that must be provided will indeed be (roughly) proportional to the clumping factor κ_c, a property that was already highlighted in Section II.3.3.1, e.g., in the course of the discussion concerning the quantiles of the queue content shown in Figs. II-12a,b.

In Figs. II-20a,b, we have plotted the P_{LP} versus ρ, for a fixed buffer size of K_q=30, a clumping factor κ_c=1,2,3,4,5, and c=1 (a) and 16 (b) respectively. As already announced in Section II.5.3, a remarkable property of the approximate approach, is that the close correspondence with the exact results remains valid for values of $\rho > 1$, although the derivations explicitly ruled out these values of the offered load, since infinite-buffer-size results were applied to deduce an expression for the pgf $L(z)$ of ℓ (e.g. (II.54)) and the approximate formulae for P_{LP} , μ_ℓ and σ_ℓ in (II.58b). In addition, these figures illustrate that P_{LP} evolves to 1 for increasing values of $\rho > 1$, as expected, and the value of c has a distinct impact on the speed of this convergence. From these results, we observe as well that, whereas the packet loss probability increases for increasing values of κ_c if $\rho < 1$, the opposite is true for $\rho > 1$. This implies that the more the packet arrival process is clumped if $\rho > 1$, the smaller the



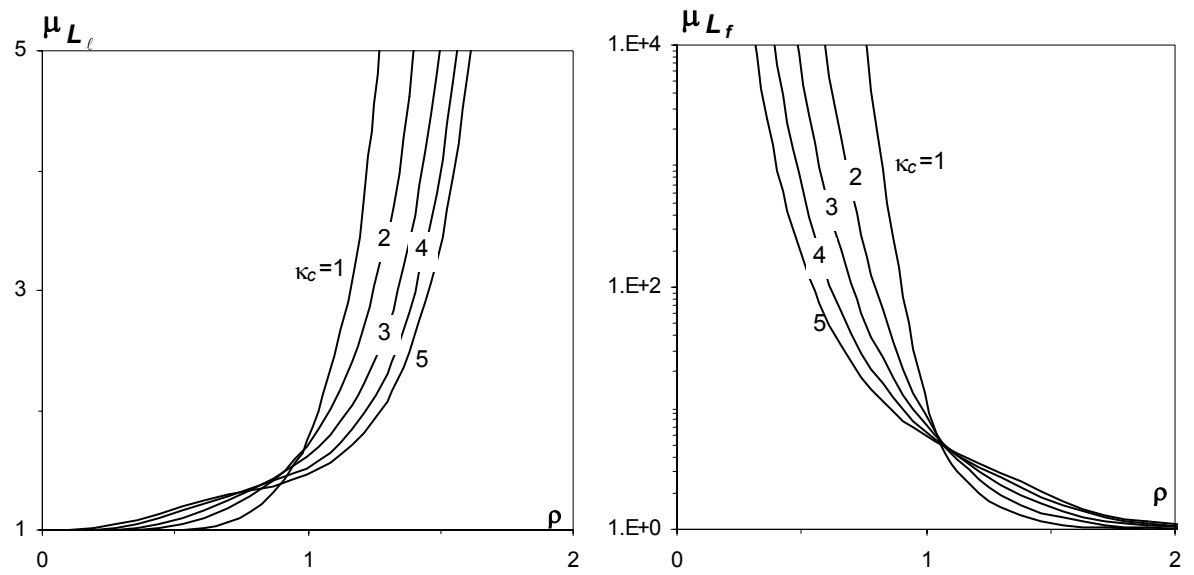
(a) mean (b) stdv

Figure II-21 : moments of the conditional packet loss process; buffer size K_q=30 and c=1

fraction of slots during which packet loss occurs, which is a somewhat counterintuitive ascertainment.

Nevertheless, Figs. II-21a,b illustrate that as far as the mean (a) and stdv (b) of the conditional packet loss process is concerned, the conventional wisdom concerning packet loss remains valid : the higher the variability of the packet arrival process (and hence, the value of κ_c), the more packets will be rejected, and the higher the variability of this (conditional) loss process. It is not difficult to check that this is also true for the unconditional loss process; note for instance that for ever increasing values of ρ, $P_{LP} \rightarrow 1$, in which case the unconditional loss moments become proportional to the conditional ones, as can be easily deduced from (II.59). These results also appear to suggest that the conditional loss process is not a highly variable one, since the corresponding coefficient of variation satisfies $\sigma_{\ell_c} / \mu_{\ell_c} < 1$, unless ρ and κ_c become extremely high, in which case this quantity can slightly exceed 1. Finally, Figs. II-21a,b once more highlight the accuracy of the approximate formulae (II.58b) for the whole range of possible values for ρ.

To conclude, we depict the mean lengths of the loss (a) and loss-free (b) periods versus ρ in Figs. II-22a,b, for K_q=30, c=16, and values of the clumping factor κ_c=1,2,3,4,5 respectively. These results, not surprisingly, indicate that, the higher the value of the offered load, the longer the loss periods and the shorter the lengths of the loss-free periods (on average) will be. From these figures we can apparently conclude that, for low values of ρ, an increase of the clumping factor yields (on average) longer loss periods and shorter loss-free periods, while for high values of ρ we observe the opposite behaviour; the tilting point of this



(a) loss (b) loss-free

Figure II-22 : mean lengths of the loss and loss-free periods versus ρ; buffer size K_q=30

behaviour is situated in the neighbourhood of the value ρ=1 for these particular values of the system parameters, but is not (necessarily) equal to it. If we focus on the loss periods, then it follows from Section II.5.4 that a loss period is terminated if the number of packet arrivals is sufficiently low (i.e., less than or equal to c); apparently the probability that this event occurs decreases for increasing κ_c if ρ is sufficiently low, leading to loss periods that typically become longer under these conditions. The reason why this behaviour is inversed when ρ increases has to do with the details and intricacies of the packet arrival process under study (i.e., the probability of having no more than c packet arrivals) and therefore does not lend itself to an easy intuitive explanation. The opposite argument holds as far as the lengths of the loss-free periods are concerned, since these are terminated if the number of new packet arrivals exceeds a certain threshold; this explains why the mean lengths of the loss and loss-free periods exhibit the opposite behaviour as ρ and κ_c increase.

II.6 Conclusions and related work

In this chapter, we have mainly focussed on a multi-server queuing model with single-slot packet transmission times, fed by N ($\geq$1) input processes that were, *independent*, and, in most cases, *identically distributed* from slot-to-slot. Initially considering the case of an infinite-capacity buffer, we have investigated quantities such as the queue and system content, and the packet waiting time and delay, for which (semi-)analytic closed-form formulae for the mean and variance, and the asymptotic tail behaviour of the pmf and ccdf, were established. The moments of these quantities were expressed, to the largest possible extent, in terms of the

'traffic parameters', being the (central) moments of the number of packet arrivals per slot, whereas the calculation of their tail distribution mandates the numerical computation of the dominant pole of their steady-state pgf, and the associated coefficient that stems from the residue theorem. In the multi-server case $c>1$, the evaluation of these performance indices also requires the numerical computation of c-1 zeroes inside the complex unit circle of the function $z^c - E(z)$, and we proposed an accurate and efficient way by which this can be avoided altogether. Finally, we also devoted some attention to the finite-buffer case, for which the loss process was investigated in considerable detail, and we managed to establish efficient computational algorithms for the main performance measures as well.

Moreover, as a supplement to the i.i.d. packet arrival process scenario, the 'clumped' packet arrival process provided a small case study that can be interpreted as a situation where packets are no longer generated in an *identical* way on a slot-per-slot basis, resulting in a highly variable process, as illustrated by many numerical examples.

The type of models that were studied have been, and are still, widely used in the performance assessment of specific subsystems that occur in telecommunication networks, such as an output buffer in a multipath self-routing switching module, which was the primary motivation at the onset of this research. Nonetheless, there are many aspects concerning these switching systems that have not been touched upon in the course of this chapter. Just to mention a few :

- in an actual switch, the 'buffering' functionality is often implemented as a shared buffer memory. Exact shared-buffer results however are notoriously hard to come by, as exemplified by [144], where the analysis of a 2×2 switch with a symmetric Bernoulli arrival process on both inlets is carried out in full. This explains why many authors resort to approximate convolution techniques ([48], [182]), which provides a worst-case scenario for the switching system. Note however that this approach provides exact results in case of a Poisson packet arrival process with independent routing;
- a switching module such as the one (briefly) described in Section II.1 is usually one of many building blocks of a switching fabric in a network node. Moreover, a packet stream typically traverses multiple network nodes from source to destination, and this explains why one is also interested in results that pertain to the end-to-end performance characteristics, for instance with respect to the packet waiting time or delay. One can often assume, and not without good reason, that the individual entities in the concatenation of consecutive output buffers that are visited by a packet (stream), operate in a more or less independent way, in which case a convolution approach can be adopted as well to generate end-to-end performance results; e.g. [3];

- in case of multi-server switching modules, it is not necessarily so that packets that belong to the same stream follow the same route, in which case a resequencing, or reassembly, unit must be provided at the receiver end. This issue has been touched upon in ([58], [65]);
- in some switching architectures, a packet (i.e., ATM cell) is not stored and forwarded as a whole, but is chopped into smaller entities in order to fit the internal data representation of the switch. This topic is looked into in [33];
- when a packet enters a switching node, it will have to undergo a number of hardware operations (such as address lookup, routing and forwarding; segmentation due to the internal data representation, …), implying that it can not be immediately forwarded to the egress of the switch, but has to reside in the memory for a minimum amount of time. Contribution [40] provides an analysis where this minimal 'hardware delay' is taken into consideration;
- the assumption of having an i.i.d. packet arrival process in an output buffer of a switch is sometimes too restrictive, depending on the architecture, its interactions with its environment, and the type of services that are being carried in the switching node. An analysis of an output buffer in a switching module with a correlated packet arrival process with independent (uniform) routing can for instance be found in [10], while a case of non-independent routing was considered in [137]. Extending the output buffer analysis to a general framework of non-i.i.d. arrival processes is exactly the focus of the next chapter.

Chapter III

Multi-server buffers with correlated arrival processes

III.1 Preface

The queuing models – and solution techniques – presented and discussed in the previous chapter focused on models that originated from the performance assessment of ATM switching elements fed by N sources, which possess the property of generating *independent* packet arrivals in an output buffer, meaning that the (total) number of packets that are generated during consecutive slots form a series of mutually statistically independent discrete random variables. This assumption, however, may not always be very realistic in an actual telecommunications environment, and it is not difficult to envision scenarios where (some) dependence, or correlation, exists between packet arrivals during subsequent slots. This dependence can be an inherent characteristic of the type of applications that are carried by the network, for instance when a number of packets that are subsequently generated on an inlet belong to the same 'session' and therefore enter the node in a grouped, or *bursty*, pattern and – on top of this – have the same destination in common as well. Hence, these are two potential sources of correlation between successive packet arrivals in an output buffer. Or, this bursty nature of the packet arrival stream may be induced by certain network conditions, for instance when a – hopefully temporary – condition of overload exists in some upstream network node, causing a temporary injection of packets in the subsequent node(s) at a high rate. Whatever the case, we need to be able to incorporate this dependence between subsequent packet arrivals into our models in order to be able to assess the performance in such situations.

A popular way to introduce correlation into the packet arrival process, is to describe it by means of a so-called *Markov*-process[(1)]. Consider therefore a set of (discrete) random variables $\{X_i : i \in I\}$ where the set of indices I represents the time-parameter, which, in a discrete-time setting, equals the set of integers $\mathbb{N}$; it is also assumed that the X_i's all have the

[(1)] Some interesting reading on the history of probability theory, in connection to A.A. Markov, can be found in [126]

same sample space Ω in common. The set of drvs $\{X_n : n \in \mathbb{N}\}$ is said to constitute a *Markov process* if it possesses the *Markov property* : let m be an integer (with $m \geq 1$) and consider any set of indices satisfying $\{n_m \in \mathbb{N} : m < m' \Leftrightarrow n_m < n_{m'}\}$, then following identity always holds

$$\Pr\left[X_{n_m} = i_{n_m} \middle| X_{n_{m-1}} = i_{n_{m-1}}, X_{n_{m-2}} = i_{n_{m-2}}, \ldots, X_{n_1} = i_{n_1}\right] = \Pr\left[X_{n_m} = i_{n_m} \middle| X_{n_{m-1}} = i_{n_{m-1}}\right] \quad ,$$

i.e., the conditional probability of finding the process in any state at any given time instant, given a series of states of the process at preceding points in time, only depends on the last known state of the process. As an example, although this was not explicitly mentioned, the stochastic process $\{s_n : n \in \mathbb{N}\}$ (or $\{q_n : n \in \mathbb{N}\}$) that describes the system (or queue) content at consecutive slot boundaries of the queuing models that were studied in the previous chapter, satisfies the Markov property, in view of the system equations (I.1a,b) (or (I.2) in case of a finite-capacity buffer) and the assumption of having i.i.d. packet arrivals.

Furthermore, when the sample – or, state – space Ω is countable, then the corresponding (discrete-time) Markov process is called a *Markov chain*. For the purpose of capturing dependence between the numbers of packet arrivals during successive slots, we will confine ourselves to the class of *time-homogeneous* and *irreducible* Markov chains with a *finite state space*. An irreducible Markov chain is subject to the property that any state can be visited from any given initial state, while a time-homogeneous Markov chain can be described by a single set of one-step transition probabilities at every time instant (i.e., slot boundary). For further details and theory on Markov chains and processes, we refer to e.g. [164].

III.2 The homogeneous packet arrival process

III.2.1 single-source modelling

We study, as before, an output buffer with N 'inlets' (i.e., sources that generate packet arrivals), and focus on any particular inlet, say inlet i. Let us adopt the irreducible and time-homogeneous Markov chain $\{b_n(i) \mid n \geq 0\}$ with state space $\Omega_{b(i)} = \{S_l : 1 \leq l \leq L\}$ to 'steer' the packet arrival process on input link i, $1 \leq i \leq N$, of an output buffer. The idea is to let different states correspond with different arrival intensities, and/or different scene lengths, etc… in the packet arrival pattern. For our current purpose, it suffices to confine ourselves to considering *identical* and *independent* stochastic arrival processes on each of the inlets. Since this set of random variables, by assumption, forms a time-homogeneous Markov chain, its evolution in time is entirely captured by the set of (one-step) *transition probabilities* p_{lm}

$$p_{lm} \triangleq \Pr[b_n(i) = S_m \mid b_{n-1}(i) = S_l] \quad , \quad 1 \leq l, m \leq L \quad ,$$

which represent the (conditional) probabilities that the arrival process (on inlet i) transits from

state S_l to state S_m at the end of a slot during which it was in state S_l; these quantities must of course satisfy the normalisation conditions

$$\sum_{m=1}^{L} p_{lm} = 1 \quad , \quad 1 \le l \le L \quad .$$

The L states of the Markov process, and the corresponding transition probabilities, are depicted in Fig. III-1.

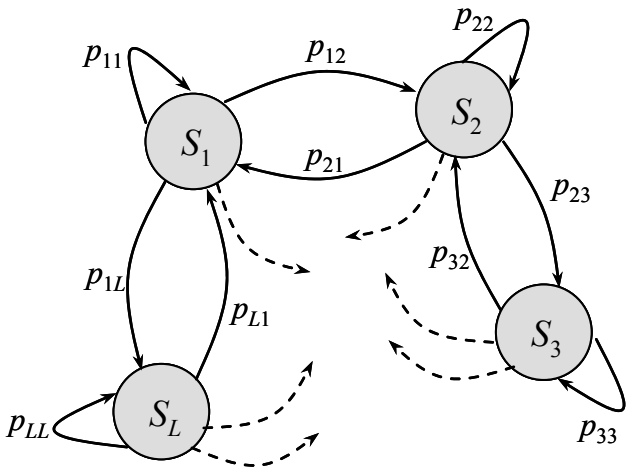


Figure III-1 State transition diagram of a time-homogeneous irreducible Markov chain

We use the stochastic process $\{b_n(i) \mid n \ge 0\}$ to *modulate* the number of packet arrivals during subsequent slots, by assuming that the number of packet arrivals during slot n only depends on the values of $b_n(i)$ and $b_{n-1}(i)$, but not on $b_{n-2}(i)$, $b_{n-3}(i)$, ⋯ (provided that $b_n(i)$ and $b_{n-1}(i)$ are known). If we then define the drvs $e_n(i)$ as the number of packet arrivals on inlet i during slot n, then these quantities can be fully characterised by defining the conditional pgfs, also called *batch-generating functions* (bgfs)

$$G_{lm}(z) \triangleq \mathbf{E}\left[z^{e_n(i)} \middle| b_n(i) = S_m, b_{n-1}(i) = S_l \right] \quad ; \quad 1 \le l, m \le L \quad .$$

Note that, in view of the previous definitions and assumptions, these functions depend on the actual values of neither n nor i, due to the Markovian and time-homogeneous nature of the modulating process, and the modelling assumption of using identical stochastic processes for generating packets on each of the inlets. In this work, we will focus on the case where the radii of convergence of the bgfs $G_{lm}(z)$, $1 \le l, m \le L$, are all larger than 1, implying that each of these functions are arbitrarily differentiable at $z=1$, and, consequently, that $e_n(i)$ has finite moments, whatever the state of the underlying Markov chain during the current and previous slot(s). In the remainder, we let $\mathcal{R}$ represent the radius of convergence of the pgm $\mathbf{Q}(z)$, as defined in Appendix D.

Let us proceed by defining random variables $a_{l,n}(i)$, $1 \le l \le L$, as

$$a_{l,n}(i) = I\big(b_n(i) = S_l\big) \quad , \quad 1 \le l \le L$$

where $I(\cdot)$ represents the *indicator* function, that equals 1 if the argument is true, and 0 otherwise. Hence, $a_{l,n}(i)=1$ only if the modulating Markov chain on inlet i is in state S_l during slot n; otherwise this drv equals 0. Evidently, $a_{l,n}(i)=1$ implies that $a_{k,n}(i)=0$ for $1 \le k \le L$ and $k \ne l$, and the set $\{\{a_{l,n}(i) : 1 \le l \le L\} : n \ge 0\}$ constitutes an alternative – and equivalent – representation of the modulating Markov chain, with identical transition probabilities, i.e.,

$$\Pr\big[a_{m,n}(i)=1\big] = \sum_{l=1}^{L} p_{lm} \Pr\big[a_{l,n-1}(i)=1\big] \;;\; G_{lm}(z) = \Big[z^{e_n(i)} \Big| a_{l,n-1}(i)=1, a_{m,n}(i)=1\Big] \quad .$$

Due to the foregoing definitions, formulas and assumptions, we may now write

$$\begin{aligned}
\sum_{j=0}^{\infty}\sum_{m=1}^{L} z^j x_m \Pr[e_n(i)=j, b_n(i)=S_m] &= \sum_{j=0}^{\infty}\sum_{m=1}^{L} z^j \left(\prod_{k=1}^{L} x_k^{a_{k,n}(i)}\right) \Pr\big[e_n(i)=j, a_{m,n}(i)=1\big] \\
&= \sum_{l=1}^{L}\sum_{m=1}^{L}\sum_{j=0}^{\infty} z^j x_m \Pr\big[e_n(i)=j, a_{l,n-1}(i)=1, a_{m,n}(i)=1\big] \\
&= \sum_{l=1}^{L}\left(\sum_{m=1}^{L} p_{lm} G_{lm}(z) x_m\right) \Pr\big[a_{l,n-1}(i)=1\big] \\
&= \sum_{l=1}^{L}\left(\prod_{k=1}^{L}\left(\sum_{m=1}^{L} p_{km} G_{km}(z) x_m\right)^{a_{k,n-1}(i)}\right) \Pr\big[a_{l,n-1}(i)=1\big] \; .
\end{aligned}$$

In view of the latter expression in the right-hand side, where the sum over m can be interpreted as being the result of a matrix product, it makes sense to define the $L \times L$ matrix $\mathbf{Q}(z)$ as

$$\mathbf{Q}(z) \triangleq \big[q_{ij}(z)\big] = \begin{bmatrix} p_{11}G_{11}(z) & p_{12}G_{12}(z) & \cdots & p_{1L}G_{1L}(z) \\ p_{21}G_{21}(z) & p_{22}G_{22}(z) & \cdots & p_{2L}G_{2L}(z) \\ \vdots & \vdots & \ddots & \vdots \\ p_{L1}G_{L1}(z) & p_{L2}G_{L2}(z) & \cdots & p_{LL}G_{LL}(z) \end{bmatrix} \quad . \tag{III.1}$$

This is the so-called *probability generating matrix* (pgm), which completely describes the packet arrival process on an input link, and that will play an important role in the analyses of the queuing models that follow hereafter.

If we now denote the $L \times 1$ (column) vectors as $\bar{\boldsymbol{x}} \equiv [x_1 \cdots x_L]^T$ and $\bar{\boldsymbol{a}}_n(i) \equiv [a_{1,n}(i) \cdots a_{L,n}(i)]^T$ (whereby $[\cdot]^T$ is the matrix-transponent of $[\cdot]$), and for any pair of $L \times 1$ vectors $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ adopt the shorthand vector notation

$$\overline{\boldsymbol{x}}^{\overline{\boldsymbol{y}}} \triangleq \prod_{l=1}^{L} x_l^{y_l} \quad , \tag{III.2}$$

then the above expression that relates the state of the Markov chain and the number of packet arrivals on inlet i during consecutive slots, can be converted into a relation between the joint pgf for the set of L+1 drvs ($e_n(i)$, $\overline{\boldsymbol{a}}_n(i)$), and the joint pgf of $\overline{\boldsymbol{a}}_{n-1}(i)$

$$\mathbf{E}\left[z^{e_n(i)}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n(i)}\right] = \mathbf{E}\left[\left(\mathbf{Q}(z)\overline{\boldsymbol{x}}\right)^{\overline{\boldsymbol{a}}_{n-1}(i)}\right] \quad . \tag{III.3}$$

Let us proceed by defining $A_i(\overline{\boldsymbol{x}})$ as the steady-state joint pgf that describes the state of the Markov chain during an arbitrary slot on a single input link,

$$A_i(\overline{\boldsymbol{x}}) \triangleq \lim_{n\to\infty} \mathbf{E}\left[\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n(i)}\right] \quad .$$

Under steady-state conditions, it becomes clear from (III.3) that this function must satisfy

$$A_i(\overline{\boldsymbol{x}}) = A_i(\mathbf{Q}(1)\overline{\boldsymbol{x}}) \quad .$$

If we signify by $\overline{\boldsymbol{\pi}} \equiv [\pi_1 \cdots \pi_L]^T$ the $L\times 1$ vector that is implicitly defined by

$$\overline{\boldsymbol{\pi}}^T = \overline{\boldsymbol{\pi}}^T \mathbf{Q}(1) \; ; \; \overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{I}} = \overline{\boldsymbol{I}} \quad , \tag{III.4a}$$

with $\overline{\boldsymbol{I}}$ the $L\times 1$ vector with all entries equal to 1, then it is easy to check that the joint pgf

$$A_i(\overline{\boldsymbol{x}}) = \overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{x}} \quad , \tag{III.4b}$$

indeed satisfies the above equation for $A_i(\overline{\boldsymbol{x}})$.

Equation (III.4a) implies that the steady-state probabilities π_l that an inlet resides in state S_l during an arbitrary slot, can be computed by solving the set of linear equations

$$\sum_{l=1}^{L} \pi_l p_{lm} = \pi_m \; ; \; 1 \le m \le L \quad ,$$

which, combined with the normalisation condition for these quantities (i.e., $\overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{I}} = \overline{\boldsymbol{I}}$) comprises the standard procedure for calculating these quantities. Since the Markov chain under consideration is irreducible and aperiodic, this set of linear equations has a unique solution. This vector is the so-called *stationary vector* of the (underlying) Markov chain.

From (III.3), we also deduce that the pgf $E_i(z)$, which describes the number of packet arrivals during an arbitrary slot on a single input link, can be written as

$$E_i(z) = \lim_{n\to\infty} \mathbf{E}\left[z^{e_n(i)}\right] = A_i\left(\mathbf{Q}(z)\overline{\boldsymbol{I}}\right) = \overline{\boldsymbol{\pi}}^T \mathbf{Q}(z)\overline{\boldsymbol{I}} \quad . \tag{III.5c}$$

The arrival process presented in this section is often referred to as a *discrete batch-Markovian arrival process* (D-BMAP, see also [130], which is the discrete version of the continuous-time process considered in [218]), where the $L\times L$ stochastic matrix $\mathbf{Q}(1)$ captures the Markovian nature of the underlying modulating process, and where the bgfs $G_{lm}(z)$, : $1\leq l,m\leq L$, characterise the batch size (i.e., the number of 'simultaneous' packet arrivals) that is generated during a slot when the transition $S_l\rightarrow S_m$ occurs at the preceding slot boundary. An important and interesting special case occurs when these bgfs correspond to a Bernoulli arrival process, with a state-transition-dependent arrival rate,

$$G_{lm}(z)=1+\gamma_{lm}(z-1) \quad ; \; 1\leq l,m\leq L \; . \tag{III.6}$$

Such a packet arrival process is a so-called *Markov modulated Bernoulli process* (MMBP), and will be the one on which we primarily focus in the remainder of this chapter, in particular when we illustrate our results with numerical examples.

Throughout the subsequent sections, it will become obvious that the decomposition of the pgm $\mathbf{Q}(z)$ (and integer powers of this matrix) into a matrix product of its associated eigenvalues and -vectors, plays an important role in our analysis. We therefore define $\mathbf{\Lambda}(z)$ as the $L\times L$ diagonal matrix containing the eigenvalues $\lambda_i(z)$, $1\leq i\leq L$, of $\mathbf{Q}(z)$, $\mathbf{W}(z)$ as the $L\times L$ matrix containing the respective left–eigenvectors $\overline{\boldsymbol{w}}_i(z)$ with eigenvalue $\lambda_i(z)$ of $\mathbf{Q}(z)$, and $\mathbf{U}(z)$ as the inverse matrix of $\mathbf{W}(z)$, meaning that these matrices satisfy the following pair of equations :

$$\begin{cases} \mathbf{W}(z)\mathbf{Q}(z)=\mathbf{\Lambda}(z)\mathbf{W}(z) \\ \mathbf{U}(z)\mathbf{W}(z)=\mathbf{W}(z)\mathbf{U}(z)=\mathbf{I} \end{cases} , \tag{III.7a}$$

where $\mathbf{I}$ represents the $L\times L$ diagonal identity matrix. These matrix equations imply that the relations

$$\begin{cases} \mathbf{Q}(z)\mathbf{U}(z)=\mathbf{U}(z)\mathbf{\Lambda}(z) \\ \mathbf{Q}(z)^k=\mathbf{U}(z)\mathbf{\Lambda}(z)^k\mathbf{W}(z) \end{cases} , \tag{III.7b}$$

hold as well, which shows that $\mathbf{U}(z)$ is the $L\times L$ matrix whose columns $\overline{\boldsymbol{u}}_i(z)$ are the right–eigenvectors, with eigenvalue $\lambda_i(z)$, of $\mathbf{Q}(z)$. Furthermore, any power of $\mathbf{Q}(z)$ can be decomposed into the matrix product that appears in the right-hand side of the second equation. The respective left- and right-eigenvectors can be calculated from these relations upon some constant factor, which can be fixed by requiring that the rows of either $\mathbf{W}(z)$ or $\mathbf{U}(z)$ are normalised

$$\mathbf{W}(z)\overline{\boldsymbol{1}}=\overline{\boldsymbol{1}} \Leftrightarrow \mathbf{U}(z)\overline{\boldsymbol{1}}=\overline{\boldsymbol{1}} \quad . \tag{III.7c}$$

From now on, we let $\lambda_1(z)$ represent the so-called *Perron-Frobenius* (PF) eigenvalue, satisfying $\lambda_1(1)=1$, while $\overline{\boldsymbol{w}}_1(z)$ and $\overline{\boldsymbol{u}}_1(z)$ are the left- and right-eigenvectors that are associated with it. In case of an irreducible and aperiodic Markov chain, the remaining eigenvalues satisfy $|\lambda_i(1)|<1$, $2\le i\le L$ (see also Appendix D).

At this point, we need to highlight the fact that, although the solution for the eigenvectors that follows from the set of equations (III.7a,b,c) works well and produces a unique and regular solution for general values of z, it unfortunately deteriorates for $z=1$. This is due to the requirement $\mathbf{W}(z)\overline{\boldsymbol{1}}=\overline{\boldsymbol{1}}$, which is rather unusual (a standard normalisation condition for the left–eigenvectors would for instance require that $\overline{\boldsymbol{w}}_i(z)\cdot\overline{\boldsymbol{w}}_i(z)^T=1$), and it can be shown (e.g. Appendix D) that the solution for $z=1$ degenerates into

$$\begin{cases}\overline{\boldsymbol{u}}_1(1)=\overline{\boldsymbol{1}}\\ \overline{\boldsymbol{u}}_i(1)=\overline{\boldsymbol{0}}\end{cases},\ 2\le i\le L \quad ; \quad \begin{cases}\overline{\boldsymbol{w}}_1(1)=\overline{\boldsymbol{\pi}}^T\\ \lim\limits_{z\to 1}\overline{\boldsymbol{w}}_i(z)\to\overline{\infty}^T\end{cases},\ 2\le i\le L \quad , \tag{III.7d}$$

where $\overline{\boldsymbol{0}}$ and $\overline{\infty}$ are $L\times 1$ column vectors whose L entries are equal to 0 and $\pm\infty$ respectively. It is important to emphasise however that, when we consider the limit $z\to 1$, then this 'degenerate' solution *still satisfies each of the equations* in (III.7a,b,c), which will be relied upon in the course of the derivations that are presented in this and the next chapter. Moreover, if we denote by $u_{ij}(z)$ the entries of $\mathbf{U}(z)$ (and hence, the i-th component of $\overline{\boldsymbol{u}}_j(z)$), we would like to underscore that the products $u_{ij}(z)\cdot\overline{\boldsymbol{w}}_j(z)$, $1\le i,j\le L$, are indeed finite for $z=1$, and these products are, in fact, the quantities that we need in our analysis; e.g. the derivations that are presented in the subsequent sections. Note for instance that the identities

$$\begin{cases}\left(u_{ij}(z)\cdot\overline{\boldsymbol{w}}_j(z)\right)\mathbf{Q}(z)=\lambda_j(z)\left(u_{ij}(z)\cdot\overline{\boldsymbol{w}}_j(z)\right)\\ \left(u_{ij}(z)\cdot\overline{\boldsymbol{w}}_j(z)\right)\overline{\boldsymbol{1}}=u_{ij}(z)\end{cases};\ 2\le j\le L\ , \tag{III.7e}$$

show that $u_{ij}(z)\cdot\overline{\boldsymbol{w}}_j(z)$ is a left eigenvector of $\mathbf{Q}(z)$ with eigenvalue $\lambda_j(z)$, and one can check by means of numerous examples that the solution of this set of equations indeed yields a finite value at $z=1$; see also the specific results for $L=2$ given in Section D.4 of Appendix D.

The primary reason for preferring this particular alternative solution for the eigenvectors, is rooted in the observation that relations such as $\overline{\boldsymbol{w}}_i(z)\overline{\boldsymbol{1}}=1$, and the simple and attractive form of $\mathbf{U}(1)$, considerably simplifies some of the results that will be deduced in the remainder of this work, as will become clear later on. More detailed information on the computation of the eigenvalues and –vectors of $\mathbf{Q}(z)$, and their behaviour and derivatives at $z=1$, can be found in Appendix D.

III.2.2 aggregation of N sources

In our performance studies of an output buffer, with packets being generated by N (identical and independent) sources, we are interested in the aggregate arrival process, i.e., the total number of packets generated per slot by the N sources. The rationale for introducing the alternative description $\overline{\boldsymbol{a}}_n(i)$ of the underlying Markov chain now becomes clear : if we define the vector of L drvs $\overline{\boldsymbol{a}}_n \triangleq [a_{1,n} \cdots a_{L,n}]^T$, where $a_{l,n}$ represents the total number of inlets in state S_l during slot n, $1 \le l \le L$, and denote by e_n the total number of packet arrivals during this slot, then we may simply write

$$a_{l,n} = \sum_{i=1}^{N} a_{l,n}(i) \ , \ 1 \le l \le L \ ; \qquad e_n = \sum_{i=1}^{N} e_n(i) \quad .$$

Through some standard derivations, the joint pgf of these drvs can be written as

$$\mathbf{E}\left[z^{e_n} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n} \right] = \mathbf{E}\left[z^{\sum_{i=1}^{N} e_n(i)} \left(\prod_{l=1}^{L} x_l^{\sum_{i=1}^{N} a_{l,n}(i)} \right) \right] = \mathbf{E}\left[\prod_{i=1}^{N} \left(z^{e_n(i)} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n(i)} \right) \right] = \prod_{i=1}^{N} \mathbf{E}\left[z^{e_n(i)} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n(i)} \right]$$

where we have exploited the property that the sets of random variables $\{e_n(i), \overline{\boldsymbol{a}}_n(i)\}$, $1 \le i \le N$, are, by assumption, mutually independent, and identically distributed. With the aid of (III.3), we then obtain

$$\mathbf{E}\left[z^{e_n} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n} \right] = \mathbf{E}\left[\left(\mathbf{Q}(z) \overline{\boldsymbol{x}} \right)^{\overline{\boldsymbol{a}}_{n-1}} \right] \quad . \tag{III.8}$$

This expression is a shorthand matrix notation for an equation that completely captures the evolution in time of the state of the Markov chain that controls the packet arrival process, and the actual number of packet arrivals that are subsequently generated. This equation will form the basis of, and will be central to, (some of) the derivations in the remainder of this work.

The foregoing deductions reveal that the set of (vectors of) drvs $\{\overline{\boldsymbol{a}}_n \ : \ n \ge 0\}$ forms a Markov chain as well, with a state space that equals

$$\Omega_{\overline{\boldsymbol{a}}} = \left\{ [l_1 \cdots l_L]^T : l_i \ge 0 \wedge \sum_{i=1}^{L} l_i = N \right\} \quad , \tag{III.9a}$$

and the cardinality of this set is given by

$$|\Omega_{\overline{\boldsymbol{a}}}| = \binom{N+L-1}{N} \ ; \ \binom{i}{j} \triangleq \frac{i!}{j!(i-j)!} \quad , \tag{III.9b}$$

which is the amount of different states that the underlying Markov chain that determines the arrival process can visit, or, equivalently, the number of different combinations by which N sources can visit L states. The condition in the right-hand side of (III.9a), expressing that the sum of the l_i's must equal N, reflects the property that

$$\sum_{l=1}^{L} a_{l,n} = N \quad , \tag{III.9c}$$

which is the mathematical translation of the observation that, during any slot, each of the N inlets visits one of the states S_l, $1 \le l \le L$.

Finally, if we define $A(\bar{x})$ as the steady-state joint pgf that describes the state of the underlying Markov chain, then from the previous results we readily obtain

$$\begin{aligned} A(\bar{x}) &\triangleq \lim_{n\to\infty} \mathbf{E}\left[\bar{x}^{\bar{a}_n}\right] = A_i(\bar{x})^N \\ &= \left(\bar{\pi}^T \bar{x}\right)^N . \end{aligned} \tag{III.10}$$

Hence, the pgf $E(z)$ associated with the drv e, that describes the (total) number of packet arrivals during an arbitrary slot, is given by

$$E(z) \triangleq \mathbf{E}\left[z^e\right] = A\left(\mathbf{Q}(z)\bar{I}\right) = \left(\bar{\pi}^T \mathbf{Q}(z)\bar{I}\right)^N \quad . \tag{III.11a}$$

Therefore, if we denote by $\mathbf{Q}'(z) \triangleq [p_{lm} G'_{lm}(z)]$ and $\mathbf{Q}''(z) \triangleq [p_{lm} G''_{lm}(z)]$ the matrices whose entries are the first and second derivatives with respect to z of the elements of $\mathbf{Q}(z)$ respectively, then a straightforward calculation leads to the following expression for $c\rho$, the average number of packet arrivals per slot :

$$c\rho \triangleq \mu_e = E'(1) = N\bar{\pi}^T \mathbf{Q}'(1)\bar{I} = N \sum_{l=1}^{L}\sum_{m=1}^{L} \pi_l p_{lm} G'_{lm}(1) \quad . \tag{III.11b}$$

For a buffer with infinite storage capacity, c output links, and packet transmission times that equal 1 slot, the equilibrium condition will require that $\rho<1$, as in (I.4). In addition, the variance of e can by similarly derived, and is given by

$$\sigma_e^2 = N\bar{\pi}^T \left(\mathbf{Q}''(1) + \mathbf{Q}'(1) - \mathbf{Q}'(1)\bar{I}\bar{\pi}^T \mathbf{Q}'(1)\right)\bar{I} = N \sum_{l=1}^{L}\sum_{m=1}^{L} \pi_l p_{lm} G''_{lm}(1) + c\rho\left(1 - \frac{c\rho}{N}\right) \quad . \tag{III.11c}$$

Analogously, the third central moment of this drv can be calculated from (e.g. (A.9))

$$\mu_{3,e} = E'''(1) - 3\left(E'(1)-1\right)E''(1) + E'(1)\left(E'(1)-1\right)\left(2E'(1)-1\right) \quad . \tag{III.11d}$$

If the arrival process is a MMBP, where the bgfs $G_{lm}(\cdot)$ can be written as in (III.6), then $E(z)$ represents a binomial process with parameters $c\rho/N$ and N respectively, e.g., expression (II.1) of the previous chapter.

The D-BMAP, and in particular the MMBP class of packet arrival processes, has been used on many occasions to model 'real' traffic streams, such as packetised voice and video traffic, as illustrated by [121], [123], [132], [173], [179], [208], [224], [229], [253], [273], and many more. Before resorting to the analysis of an output buffer, with N sources that generate packets according to the D-BMAP described in this section, let us first briefly examine some of its features in more detail.

III.2.3 (in)dependence and correlation

The question that we want to address is the following : under which circumstances does this D-BMAP arrival process produce an i.i.d. packet arrival pattern ? Since the packet arrival process is the aggregation of N identical and independent processes, it will generate an i.i.d. arrival pattern during consecutive slots, if the packet arrival process on a single input link does so. This will be the case if the steady-state joint pgf of the number of packet arrivals generated by a single arrival stream during two consecutive slots equals the product of the two corresponding (steady-state) marginal pgfs. If we denote by $(\overline{\boldsymbol{a}}(i), e(i))$, and $(\overline{\boldsymbol{a}}_+(i), e_+(i))$ and $(\overline{\boldsymbol{a}}_-(i), e_-(i))$ the state of the arrival process and the number of packet arrivals during an arbitrary slot, and the following and preceding slot respectively, on the i-th input link, then with the help of (III.3) we can derive that

$$\mathbf{E}\left[y^{e(i)} z^{e_+(i)} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_+(i)}\right] = \mathbf{E}\left[y^{e(i)} (\mathbf{Q}(z)\overline{\boldsymbol{x}})^{\overline{\boldsymbol{a}}(i)}\right] = \mathbf{E}\left[(\mathbf{Q}(y)\mathbf{Q}(z)\overline{\boldsymbol{x}})^{\overline{\boldsymbol{a}}_-(i)}\right] = A_i(\mathbf{Q}(y)\mathbf{Q}(z)\overline{\boldsymbol{x}})$$
$$= \overline{\boldsymbol{\pi}}^T \mathbf{Q}(y)\mathbf{Q}(z)\overline{\boldsymbol{x}} \quad .$$

The D-BMAP will generate an i.i.d. arrival pattern if and only if

$$\mathbf{E}\left[y^{e(i)} z^{e_+(i)}\right] = \mathbf{E}\left[y^{e(i)}\right]\mathbf{E}\left[z^{e_+(i)}\right] \quad ,$$

leading to the matrix equation

$$\left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}(y)\mathbf{Q}(z)\overline{\boldsymbol{I}}\right) = \left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}(y)\overline{\boldsymbol{I}}\right)\left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}(z)\overline{\boldsymbol{I}}\right) \quad .$$

Since this condition must be satisfied for all values of y and z, we can distinguish two cases : on the one hand, it will be fulfilled if

$$\mathbf{Q}(z)\overline{\boldsymbol{I}} = \overline{\boldsymbol{I}}\,\overline{\boldsymbol{\pi}}^T \mathbf{Q}(z)\overline{\boldsymbol{I}} = \overline{\boldsymbol{I}}\, E_i(z) \quad ,$$

leading to the set of equations

$$\sum_{m=1}^{L} p_{lm} G_{lm}(z) = E_i(z) \ , \ \forall\, 1 \le l \le L \quad . \tag{III.12a}$$

On the other hand, the condition for independence is satisfied as well if

$$\overline{\boldsymbol{\pi}}^T \mathbf{Q}(y) = \overline{\boldsymbol{\pi}}^T \mathbf{Q}(y) \overline{\boldsymbol{I}}\, \overline{\boldsymbol{\pi}}^T = E_i(y) \overline{\boldsymbol{\pi}}^T \quad .$$

which, in turn, yields the following set of equations

$$\sum_{l=1}^{L} \pi_l p_{lm} G_{lm}(y) = \pi_m E_i(y) \ , \ \forall\, 1 \le m \le L \quad . \tag{III.12b}$$

Hence, if the bgfs $G_{lm}(z)$ are such that they satisfy either one of the set of equations (III.12a,b), then the corresponding packet arrival process will be an i.i.d. process. Observe that (III.12a,b) are automatically fulfilled for $z=1$ and $y=1$ respectively. A closer inspection of these two sets of equations reveals that (III.12a) corresponds to the requirement that, for any value of $z \in \mathbb{C}$ (with $|z|<\mathcal{R}$), the $L \times 1$ vector $\overline{\boldsymbol{I}}$ is a right-eigenvector of $\mathbf{Q}(z)$ with eigenvalue $E_i(z)$. Similarly, (III.12b) comprises the property that the stationary vector $\overline{\boldsymbol{\pi}}^T$ is a left-eigenvector of the pgm $\mathbf{Q}(z)$, with eigenvalue $E_i(z)$ as well. Hence, either way, $E_i(z)$ will be an eigenvalue of $\mathbf{Q}(z)$ if this pgm represents an i.i.d. process, and since $E_i(1)=1$, we can only conclude that this pgf is the PF-eigenvalue of $\mathbf{Q}(z)$. Consequently, these observations enable us to put forward the premise that

if the packet arrival stream that is generated by a D-BMAP corresponds to an i.i.d. process, then the PF-eigenvalue of the associated pgm $\mathbf{Q}(z)$ *will be equal to the pgf* $E_i(z)$ *that describes the number of packet arrivals on an inlet during an arbitrary slot.*

Moreover, it is not difficult to verify that (III.12a) and (III.12b) do not necessarily lead to an equivalent set of equations. In order to exemplify this notion, consider for instance the 2×2 pgm

$$\mathbf{Q}(z) = \begin{bmatrix} \alpha G_1(z) & (1-\alpha) G_2(z) \\ \alpha G_1(z) & (1-\alpha) G_2(z) \end{bmatrix} \ ; \ \begin{matrix} \overline{\boldsymbol{\pi}}^T = [\alpha \ \ (1-\alpha)] \\ E_i(z) = \alpha G_1(z) + (1-\alpha) G_2(z) \end{matrix} \ , \tag{III.13a}$$

which satisfies (III.12a) but not (III.12b) for arbitrary bgfs $G_1(z)$ and $G_2(z)$, while the opposite is true for the pgm

$$\mathbf{Q}(z) = \begin{bmatrix} \alpha G_1(z) & (1-\alpha) G_1(z) \\ \alpha G_2(z) & (1-\alpha) G_2(z) \end{bmatrix} \ ; \ \begin{matrix} \overline{\boldsymbol{\pi}}^T = [\alpha \ \ (1-\alpha)] \\ E_i(z) = \alpha G_1(z) + (1-\alpha) G_2(z) \end{matrix} \ . \tag{III.13b}$$

Let us now also take a brief look at the criterion for having an (*un*)*correlated* packet arrival process. From the previous results, it is not difficult to deduce that the coefficient of correlation of the pair of drvs ($e(i)$, $e_+(i)$), under steady-state conditions, satisfies

$$\rho_{e(i),e_+(i)} \triangleq \frac{\mathbf{E}[e(i)e_+(i)] - \mu_{e(i)}^2}{\sigma_{e(i)}^2} = \frac{\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)^2 \overline{\boldsymbol{I}} - \left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1) \overline{\boldsymbol{I}}\right)^2}{\overline{\boldsymbol{\pi}}^T \left(\mathbf{Q}''(1) + \mathbf{Q}'(1) - \mathbf{Q}'(1) \overline{\boldsymbol{I}} \overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)\right) \overline{\boldsymbol{I}}} = \frac{\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)^2 \overline{\boldsymbol{I}} - \left(\frac{c\rho}{N}\right)^2}{\overline{\boldsymbol{\pi}}^T \mathbf{Q}''(1) \overline{\boldsymbol{I}} + \left(\frac{c\rho}{N}\right) - \left(\frac{c\rho}{N}\right)^2} \quad ,$$

and one can verify that this quantity also represents the coefficient of correlation of the *aggregate* packet arrival process. Consequently, the (aggregate) packet arrival process will be uncorrelated if the condition

$$\left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)^2 \overline{\boldsymbol{I}}\right) = \left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1) \overline{\boldsymbol{I}}\right)\left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1) \overline{\boldsymbol{I}}\right) \quad ,$$

is fulfilled. Obviously, this is a much weaker condition than the one for independence : an i.i.d. arrival process is perforce uncorrelated, while the inverse is not necessarily true. The previous matrix equation will hold for all parameter sets that satisfy

$$\sum_{k=1}^{L} \sum_{l=1}^{L} \sum_{m=1}^{L} \pi_k p_{kl} G'_{kl}(1) p_{lm} G'_{lm}(1) = \left(\frac{c\rho}{N}\right)^2 \quad , \tag{III.14}$$

and one can easily ensure oneself of the fact that the two solutions (III.12a,b) of the matrix condition for having independent arrival processes indeed satisfy this relation.

Let us conclude by considering an MMBP packet arrival process as a small example, whereby the bgfs that characterise the batch sizes satisfy the generic form (III.6). Then the pgm $\mathbf{Q}(z)$ that entirely captures the packet arrival process, contains $L(2L\text{-}1)$ parameters that can be chosen independently. Requiring that the arrival process is i.i.d. then imposes either one of the L (non-linear) conditions (III.12a,b) on this set of $L(2L\text{-}1)$ source parameters, while the above condition for having an uncorrelated arrival process imposes just one single condition on these parameters. Hence, in all probability, many parameter sets for the packet arrival process can be composed that generate an independent and/or uncorrelated process.

III.2.4 the burst factor

The coefficient of correlation $\rho_{e(i),e_+(i)}$ can, in principle, be applied as a benchmark for the *variability*, or *burstiness*, of the D-BMAP under consideration. One drawback that this quantity suffers from however, is that it lacks a clear probabilistic interpretation in terms of the traffic parameters that determine the arrival process (i.e., what does a particular value for this quantity, say 0.5, exactly mean ?). We therefore prefer a somewhat alternative approach

to tackle the issue of identifying suitable parameters to capture the burstiness of the (aggregate) packet arrival process.

Let us commence by pointing out that, as before, μ_e, σ_e^2 and $\mu_{3,e}$ represent the mean, variance and third central moment of the total number of packet arrivals during an arbitrary slot, that can be computed from (III.11b-d). These are also the moments of the number of packet arrivals during any slot, if the arrival process were an i.i.d. process described by the pgf $E(z)$ given by (III.11a). In addition, we define e_M, with steady-state pgf $E_M(z)$, as the drv that represents the *total number of packet arrivals during M consecutive slots*. Then, applying the results of Section III.2.2, one can show that

$$E_M(z) \triangleq \lim_{n\to\infty} \mathbf{E}\left[z^{e_{n-1}+e_{n-2}+\cdots+e_{n-M}}\right] = \left(\overline{\boldsymbol{\pi}}^T \mathbf{Q}(z)^M \overline{\boldsymbol{I}}\right)^N \quad .$$

Making use of decomposition (III.7b) of $\mathbf{Q}(z)^M$ and $\mathbf{W}(z)\overline{\boldsymbol{I}} = \overline{\boldsymbol{I}}$, we can establish the relation

$$E_M(z) = \left(\overline{\boldsymbol{\pi}}^T \mathbf{U}(z)\boldsymbol{\Lambda}(z)^M \overline{\boldsymbol{I}}\right)^N \quad ,$$

from which we can deduce that the variance of e_M is given by

$$\begin{aligned}\mathcal{V}[E_M] = N\overline{\boldsymbol{\pi}}^T \Big(&\mathbf{U}''(1)\boldsymbol{\Lambda}(1)^M + 2M\mathbf{U}'(1)\boldsymbol{\Lambda}'(1)\boldsymbol{\Lambda}(1) + M\mathbf{U}(1)\boldsymbol{\Lambda}''(1)\boldsymbol{\Lambda}(1) \\ &+ M(M-1)\mathbf{U}(1)\boldsymbol{\Lambda}'(1)^2\Big)\boldsymbol{\Lambda}(1)^{M-2}\overline{\boldsymbol{I}} \\ + N\overline{\boldsymbol{\pi}}^T &\left(\mathbf{U}'(1)\boldsymbol{\Lambda}(1) + M\mathbf{U}(1)\boldsymbol{\Lambda}'(1)\right)\boldsymbol{\Lambda}(1)^{M-1}\overline{\boldsymbol{I}}\left\{1 - \overline{\boldsymbol{\pi}}^T\left(\mathbf{U}'(1)\boldsymbol{\Lambda}(1) + M\mathbf{U}(1)\boldsymbol{\Lambda}'(1)\right)\boldsymbol{\Lambda}(1)^{M-1}\overline{\boldsymbol{I}}\right\} \quad .\end{aligned}$$

Since the non-PF eigenvalues satisfy $|\lambda_i(1)|<1$ for $2\le i\le L$, the following identities hold

$$\lambda_1(1)^M = 1 \ , \ \forall\ M \ge 0 \qquad ; \lim_{M\to\infty} \lambda_i(1)^M = 0 \ , \ 2 \le i \le L \ .$$

We then define, taking into account that $\mathbf{E}[e_M] = M\mu_e = Mc\rho$,

$$\kappa_b \triangleq \lim_{M\to\infty} \frac{\mathcal{V}[E_M]}{M\mathcal{V}[E]} = \lim_{M\to\infty} \frac{\mathbf{E}\left[\left(e_M - M\mu_e\right)^2\right]}{M\sigma_e^2} \quad , \qquad \text{(III.15a)}$$

implying that this quantity equals the variance of the total number of packet arrivals during a 'very large' time interval, relative to a scenario with i.i.d. packet arrivals described by the pgf $E(z)$. Therefore, it can be interpreted as being a measure for the variability, or burstiness, in the arrival process, measured over long time periods, that is induced by the correlation of the D-BMAP under consideration. In the remainder, κ_b will be referred to as the *burst factor*. The results that will be deduced in the subsequent sections will reveal that this parameter

plays a primary role in the calculation of performance indices such as the mean and variance of the queue and system content.

In view of the previous, we can show that κ_b can be calculated as

$$\kappa_b \sigma_e^2 = \lim_{M\to\infty}\left[N\sum_{i=1}^{L}\pi_i\left\{2u'_{i1}(1)\lambda'_1(1)+u_{i1}(1)\left(\lambda''_1(1)+(M-1)\lambda'_1(1)^2+\lambda'_1(1)\right)\right\}\right.$$
$$\left. -MN\left(\sum_{i=1}^{L}\pi_i u_{i1}(1)\lambda'_1(1)\right)^2 - 2N\left(\sum_{i=1}^{L}\pi_i u_{i1}(1)\lambda'_1(1)\right)\left(\sum_{i=1}^{L}\pi_i u'_{i1}(1)\right)\right] .$$

Then with the use of $\overline{\boldsymbol{\pi}}^T\overline{\boldsymbol{u}}_1(1)=1$, we easily verify that this expression for κ_b reduces to

$$\kappa_b = N\left(\lambda''_1(1)+\lambda'_1(1)-\lambda'_1(1)^2\right)\Big/\sigma_e^2 = \boldsymbol{\mathcal{V}}\left[\lambda_1^N\right]\Big/\boldsymbol{\mathcal{V}}[E] \quad . \tag{III.15b}$$

Furthermore, if we define the parameter κ_s as the third central moment of the total number of packet arrivals during a 'sufficiently long' time period, relative to the case of an i.i.d. arrival process described by the pgf $E(z)$,

$$\kappa_s \triangleq \lim_{M\to\infty}\frac{\mathbf{E}\left[\left(e_M - M\mu_e\right)^3\right]}{M\mu_{3,e}} \quad ,$$

then one can show in an analogous – albeit somewhat tedious – manner as before that this quantity can be calculated from the formula

$$\kappa_s \triangleq N\left(\lambda'''_1(1)-3\left(\lambda'_1(1)-1\right)\lambda''_1(1)+\lambda'_1(1)\left(\lambda'_1(1)-1\right)\left(2\lambda'_1(1)-1\right)\right)\Big/\mu_{3,e} \quad . \tag{III.15c}$$

Consequently, the penultimate expression that defines κ_s indicates that this quantity can be viewed as a benchmark for the *asymmetry*, or *skewness*, of the packet arrival process, measured over a long time period, relative to the case where packet arrivals are generated by an i.i.d. process with pgf $E(z)$[(2)]. When appropriate, κ_s will be referred to as the *skew factor*.

In Appendix D, we also show that $N\lambda'_1(1)=c\rho$. Hence, relying on the previous formulae, and the computational rules that are established in Appendix A for calculating the first three (central) moments of a drv from its pgf (see (A.9)), we can now venture to make the following statement : although $\lambda_1(z)^N$ is not a pgf as such, the '(central) moments' that are associated to it (i.e., the quantities that are calculated as if it were a pgf by taking the appropriate derivatives with respect to z for z=1), are equal to *the (central) moments of the*

(2) similar to the mean and variance of a sum of independent drvs, the third central moment of such a sum is equal to the sum of the third central moments, which accounts for the denominator in the definition of the skew factor

drv that represents the total number of packet arrivals that are generated during a sufficiently long time period, divided by the length of the period under consideration. This assessment has just been checked and verified for the first three (central) moments, and renders a useful probabilistic interpretation to the PF-eigenvalue, and the parameters κ_b and κ_s that are related to it. Also observe that the quantities $\kappa_b\sigma_e^2$ and $\kappa_s\mu_{3,e}$ are linear in (i.e., proportional to) the number of sources that are being multiplexed, which is one more reason why these parameters are prime traffic descriptors.

To conclude, let us return to the property that was established in the foregoing section, expressing that the PF-eigenvalue of $\mathbf{Q}(z)$ equals $E_i(z)$ in case of an i.i.d. packet arrival process, implying that $\lambda_1(z)^N \equiv E(z)$ under those circumstances. Hence, in view of the definitions of κ_b and κ_s and their relation to the derivatives of the PF-eigenvalue, we can make the following pronouncement as well : if the D-BMAP generates an i.i.d. packet stream, then the burst and skew factor both satisfy $\kappa_b=\kappa_s=1$. This is, generally speaking, not necessarily true when the process is uncorrelated, but not independent from slot-to-slot.

III.3 The buffer content in a multi-server output buffer with a D-BMAP

The buffer behaviour of discrete-time queuing models with correlated inputs and single-slot packet transmission times have been primarily studied for single-server buffers ([129], [190], [207], [243], [244], [265], [274]), and although multi-server models are rather scarce, they have been investigated on some occasions (e.g., [210], [270]); most of the derivations that follow hereafter are based on the results presented in [14]. Let us, as in the foregoing chapter, consider an output buffer with infinite storage capacity, meaning that all arriving packets are allowed to enter the buffer, where they may be temporarily stored to await their transmission. The buffer has c transmission channels via which packets are transmitted and subsequently leave the system. Considering packet transmission times of a single slot implies that up to a maximum of c packets can leave the buffer system during any slot – provided that that many packets were available for transmission at the beginning of the slot. With s_n representing the system content at the beginning of slot n – i.e., just after packet departures (if any), but before new arrivals – then the system equation established in Chapter 1, that captures the evolution of this drv during consecutive slots

$$s_{n+1} = (s_n - c)^+ + e_n \quad , \tag{I.1a}$$

is still valid under these circumstances. We assume that the equilibrium condition $\rho < 1$ is satisfied, implying that the system evolves to a stochastic equilibrium after a sufficiently long period of time.

It is now no longer the case that $\{s_n : n \in \mathbb{N}\}$ in itself constitutes a Markov chain, since the packet arrival process is no longer an i.i.d. process, and if we let packet arrivals be generated by the D-BMAP described in Section III-2, then it becomes clear that we must extend our state description with the vector $\overline{\boldsymbol{a}}_n$ that keeps track of the state of the modulating Markov chain – that steers the packet arrival process – during consecutive slots. For reasons that will be clarified later on, we prefer the pair $(s_n, \overline{\boldsymbol{a}}_{n-1})$ to describe the system state during subsequent slots, i.e., the (total) number of packets in the buffer at the beginning of slot n (including those that will be transmitted during this slot), and the state of the underlying Markov process during the *preceding* slot. Note that due to the relation (III.9c), one of the components of $\overline{\boldsymbol{a}}_{n-1}$ could be omitted from this state description. Nevertheless, we wish to conserve symmetry in the set of drvs that represents the state of the underlying Markov chain as much as possible, and therefore opt to include the complete set $\overline{\boldsymbol{a}}_{n-1}$ in the state vector.

Popular solution methods for a queuing system such as the one described above, are for instance the matrix-analytic approach ([203], [218], [227], [129], [263], [274]), and the spectral decomposition analysis technique ([207], [210], [211], [243], [244], [265], [270]). As illustrated in [14], we prefer a somewhat different approach (although related to the latter one), that enables us to express the relevant performance indices – such as mean and variance of the quantities under investigation – to the largest possible extent in terms of the traffic parameters. In addition, this will also yield a logical and natural extension of the dominant pole approximation technique for the computation of the tail of the pmf or ccdf of these quantities, as we will ascertain in Section III.3.6.

The continuation of the analysis of the system content involves deriving an expression for $P_s(z, \overline{\boldsymbol{x}})$, the steady-state joint pgf of the set of L+1 drvs $(s_n, \overline{\boldsymbol{a}}_{n-1})$

$$P_s(z, \overline{\boldsymbol{x}}) \triangleq \lim_{n \to \infty} P_{s,n}(z, \overline{\boldsymbol{x}}) \; ; \; P_{s,n}(z, \overline{\boldsymbol{x}}) \triangleq \mathbf{E}\left[z^{s_n} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_{n-1}} \right] \quad ,$$

which basically consists of two steps.

III.3.1 establishing a functional equation for $P_s(z, \overline{\boldsymbol{x}})$

Combining the above definition with system equation (I.1a) for s_{n+1} leads to

$$P_{s,n+1}(z, \overline{\boldsymbol{x}}) = \mathbf{E}\left[z^{s_{n+1}} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n} \right] = \mathbf{E}\left[z^{(s_n - c)^+ + e_n} \overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n} \right] \quad .$$

A crucial observation is that, in view of the Markovian nature of the packet arrival process, and provided that $\overline{\boldsymbol{a}}_{n-1}$ is known, neither e_n nor $\overline{\boldsymbol{a}}_n$ will depend on the actual value of s_n. This implies that we may invoke (III.8) to write

$$P_{s,n+1}(z,\bar{\boldsymbol{x}}) = \mathbf{E}\left[z^{(s_n - c)^+} \left(\mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{a}}_{n-1}} \right] \quad .$$

Applying standard probabilistic/stochastic calculation techniques, this expression for the joint pgf $P_{s,n+1}(z,\bar{\boldsymbol{x}})$ can be further converted into

$$\begin{aligned} P_{s,n+1}(z,\bar{\boldsymbol{x}}) &= \mathbf{E}\left[z^{(s_n - c)^+} \left(\mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{a}}_{n-1}} \right] \\ &= z^{-c}\mathbf{E}\left[z^{s_n} \left(\mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{a}}_{n-1}} \{ s_n \geq c \} \right] + \mathbf{E}\left[\left(\mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{a}}_{n-1}} \{ s_n < c \} \right] \\ &= z^{-c} P_{s,n}\left(z, \mathbf{Q}(z)\bar{\boldsymbol{x}}\right) + \mathbf{E}\left[\left(1 - z^{s_n}\right) \left(\mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{a}}_{n-1}} \{ s_n < c \} \right] . \end{aligned}$$

Letting $n \to \infty$ and assuming steady–state conditions, we thus obtain a *functional equation* for the joint steady–state pgf $P_s(z,\bar{\boldsymbol{x}})$

$$P_s(z,\bar{\boldsymbol{x}}) = z^{-c}\left\{ P_s\left(z, \mathbf{Q}(z)\bar{\boldsymbol{x}}\right) + R\left(z, \mathbf{Q}(z)\bar{\boldsymbol{x}}\right) \right\} \quad , \tag{III.16a}$$

with

$$\begin{aligned} R(z,\bar{\boldsymbol{x}}) &\triangleq \mathbf{E}\left[\left(z^c - z^s\right) \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}} \{ s < c \} \right] = \sum_{j=0}^{c-1} \left(z^c - z^j\right) \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} \left(\prod_{i=1}^{L} x_i^{l_i} \right) \Pr\left[s = j, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}} \right] \\ &= (z-1) \sum_{j=0}^{c-1} z^j \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{l}}} \Pr\left[s \leq j, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}} \right] , \end{aligned} \tag{III.16b}$$

where we use the same kind of notational convention as in (III.2) for $\bar{\boldsymbol{x}}^{\bar{\boldsymbol{l}}}$, with $\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}$. In these expressions, the drv s represents the system content at the beginning of an arbitrary slot, while the set of drvs $\bar{\boldsymbol{a}} = [a_1 \cdots a_L]^T$ describes the state of modulating Markov chain of the arrival process during the preceding slot; the sum for the set of integers $\bar{\boldsymbol{l}} = [l_1 \cdots l_L]^T$ in this expression runs over all possible values of $\bar{\boldsymbol{l}}$ that belong to the sample space $\Omega_{\bar{\boldsymbol{a}}}$ of $\bar{\boldsymbol{a}}$, defined by (III.9a). These equations define a functional equation for the joint steady-state pgf $P_s(z,\bar{\boldsymbol{x}})$ of the system state $(s,\bar{\boldsymbol{a}})$, that inherently contains all information concerning the buffer behaviour. Under some specific circumstances, quantities such as the mean, variance, and tail distribution of the system content can be extracted from this equation through some specific manipulations; e.g. [7], [261], [265], [266]. Nevertheless, this requires an ad hoc approach that often strongly depends on the specific details of the arrival process under consideration, and may or may not be extendable under different modelling assumptions concerning the packet arrival pattern. Instead of following these paths, we attempt to

establish a generic solution technique for this functional equation that leads to an explicit expression for $P_S(z,\bar{\boldsymbol{x}})$.

Also note that the expression for $R(z,\bar{\boldsymbol{x}})$ – generally speaking – contains $c\cdot|\Omega_{\bar{\boldsymbol{a}}}|$ unknown probabilities

$$p(j,\bar{\boldsymbol{l}}) \triangleq \frac{1}{c(1-\rho)}\Pr\left[s \le j, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}}\right] \;;\; 0 \le j \le c-1 \;,\; \bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}} \quad , \tag{III.16c}$$

that are yet to be determined. The factor $c(1-\rho)$ in the denominator in the right-hand side of this expression accounts for the normalisation of these quantities, since, via a similar calculation as in Section II.3.1 (e.g. (II.8)) by expressing that – under steady-state conditions – the average number of packet arrivals equals the average number of packet departures, one can prove that they satisfy

$$\sum_{j=0}^{c-1} \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} p(j,\bar{\boldsymbol{l}}) = 1 \quad . \tag{III.16d}$$

Throughout the subsequent sections, it will become clear how these quantities can be computed, and an adequate approximation procedure will be introduced as well.

III.3.2 solving the functional equation

By repeated substitutions of the argument $\bar{\boldsymbol{x}}$ by $\mathbf{Q}(z)\bar{\boldsymbol{x}}$ in the functional equation (III.16a), we can deduce that

$$P_S(z,\bar{\boldsymbol{x}}) = z^{-Hc} P_S\left(z,\mathbf{Q}(z)^H \bar{\boldsymbol{x}}\right) + \sum_{h=1}^{H} z^{-hc} R\left(z,\mathbf{Q}(z)^h \bar{\boldsymbol{x}}\right) \quad ,$$

for all $H \ge 1$. Due to the occurrence of different powers of $\mathbf{Q}(z)$ in this expression, it seems appropriate to decompose these matrices into products of (matrices of) their eigenvalues and – vectors. Consequently, based on relations (III.7a,b,c), $\mathbf{Q}(z)^h \bar{\boldsymbol{x}}$ can be expressed as

$$\mathbf{Q}(z)^h \bar{\boldsymbol{x}} = \mathbf{U}(z)\boldsymbol{\Lambda}(z)^h \mathbf{W}(z)\bar{\boldsymbol{x}} \quad .$$

To proceed, we first define the functions $F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)$, with $\bar{\boldsymbol{l}}, \bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}}$, as

$$\prod_{i=1}^{L}\left(\sum_{j=1}^{L} u_{ij}(z)x_j\right)^{l_i} \triangleq \sum_{\bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)\bar{\boldsymbol{x}}^{\bar{\boldsymbol{m}}} \quad \Leftrightarrow \quad (\mathbf{U}(z)\bar{\boldsymbol{x}})^{\bar{\boldsymbol{l}}} \triangleq \sum_{\bar{\boldsymbol{m}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)\bar{\boldsymbol{x}}^{\bar{\boldsymbol{m}}} \tag{III.17a}$$

(with $u_{ij}(z)$ the i–th element of the vector $\bar{\boldsymbol{u}}_j(z)$), i.e., $F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)$ is the coefficient of $\bar{\boldsymbol{x}}^{\bar{\boldsymbol{m}}}$ that appears in the expression in both sides of this equality. Observe that, due to the normalisation $\mathbf{U}(z)\bar{\boldsymbol{I}}=\bar{\boldsymbol{I}}$, these functions satisfy

$$\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)=1 \;;\; \forall\, \bar{\boldsymbol{l}}\in\Omega_{\bar{\boldsymbol{a}}} \quad . \tag{III.17b}$$

These functions consist of combinations of the components of $\mathbf{U}(z)$ (these functions are, in fact, the equivalent of the N-fold Kronecker product $\mathbf{U}(z)\otimes\mathbf{U}(z)\otimes\cdots\otimes\mathbf{U}(z)$), and deriving closed-form expressions for them may be complicated for general N and L – although a computational procedure, based on a recursion on the value of N, can be constructed. In case of L=2 for instance, these functions are given by (with $\bar{\boldsymbol{l}}\equiv(l,N-l)$ and $\bar{\boldsymbol{m}}\equiv(m,N-m)$)

$$F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)=\sum_{j=(l+m-N)^+}^{\min\{l,m\}}\binom{l}{j}\binom{N-l}{m-j}u_{11}(z)^j u_{12}(z)^{l-j}u_{21}(z)^{m-j}u_{22}(z)^{N-l-m+j}\;, \tag{III.17c}$$

Nevertheless, the results presented in the subsequent sections will reveal that we manage to establish (approximate) expressions for the main performance measures that altogether avoid the calculation of these functions.

The above definition for $F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)$ enables us to write

$$\begin{aligned}\left(\mathbf{Q}(z)^h\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{l}}} &=\left(\mathbf{U}(z)\boldsymbol{\Lambda}(z)^h\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{l}}}=\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}}F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)\prod_{i=1}^{L}\left(\lambda_i(z)^h\bar{\boldsymbol{w}}_i(z)\bar{\boldsymbol{x}}\right)^{m_i}\\ &\equiv\sum_{\bar{\boldsymbol{m}}}F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)\left(\boldsymbol{\Lambda}(z)^h\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}\;.\end{aligned}$$

For sake of completeness, we would like to point out that, in spite of the 'degenerated' solution (III.7d) for $\bar{\boldsymbol{w}}_i(1)$ and $\bar{\boldsymbol{u}}_i(1)$, the terms in the sum over $\bar{\boldsymbol{m}}$ in the right-hand side of this equality assume finite values for z=1, and any given value of $\bar{\boldsymbol{l}}$ and $\bar{\boldsymbol{m}}$. In other words, although z=1 is a singularity of the non-PF left–eigenvectors, it is a *removable singular point* of the above formula(e). In this respect, it is important to note that the sum over $\bar{\boldsymbol{m}}$ in the right-hand side of the above equation consists of (powers of) the components of $u_{ij}(z)\cdot\bar{\boldsymbol{w}}_j(z)$ (multiplied by some factor that contains the eigenvalues), $1\le i,j\le L$; which, in view of the remarks concerning (III.7d,e), explains why both hand sides of the above expression can be unambiguously evaluated at z=1.

Consequently, taking into account (III.16b,d) together with the previous relation, then the last expression for $P_S(z,\bar{\boldsymbol{x}})$ can be transformed into

$$P_S(z,\bar{\boldsymbol{x}}) = \sum_{\bar{\boldsymbol{l}}} \mathbf{E}\left[z^s \{\bar{\boldsymbol{a}} = \bar{\boldsymbol{l}}\} \right] \sum_{\bar{\boldsymbol{m}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z) z^{-Hc} \prod_{i=1}^{L} \left(\lambda_i(z)^H \, \bar{\boldsymbol{w}}_j(z) \bar{\boldsymbol{x}} \right)^{m_i}$$
$$+ c(1-\rho)(z-1) \sum_{j=0}^{c-1} z^j \sum_{\bar{\boldsymbol{l}}} p(j,\bar{\boldsymbol{l}}) \sum_{\bar{\boldsymbol{m}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z) \sum_{h=1}^{H} z^{-hc} \prod_{i=1}^{L} \left(\lambda_i(z)^h \, \bar{\boldsymbol{w}}_i(z) \bar{\boldsymbol{x}} \right)^{m_i} \quad ,$$

which holds for any value of $H \geq 1$. As before (and unless mentioned otherwise), the sums for $\bar{\boldsymbol{l}}$ and $\bar{\boldsymbol{m}}$ in these (and the following) expressions run over all $\bar{\boldsymbol{l}}, \bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}}$. If we now let H approach infinity, then the sum for h in the second term of the right-hand side will converge for those values of z for which all $|\lambda_i(z)|$ are bounded by $|\lambda_i(z)| < |z|^{c/N}$, $1 \leq i \leq L$. Such values of z exist :

LEMMA : $|\lambda_i(z)| < 1$ for all $z \in \boldsymbol{C}_1 = \{ z \in \mathbb{C} : |z| = 1 \wedge z \neq 1 \}$

PROOF :

Consider a generic pgm $\mathbf{Q}(z)$, given by (III.25), and suppose that – provided that we are in state i during a slot – the batch sizes that are generated during the following slot are multiples of $\boldsymbol{\ell}_i$ packets, $1 \leq i \leq L$. If we require that $\boldsymbol{\ell}_i = 1$, $\forall\ 1 \leq i \leq L$, then this necessarily implies that

$$\left| \sum_{j=1}^{L} p_{ij} G_{ij}(z) \right| \leq \sum_{j=1}^{L} p_{ij} \left| G_{ij}(z) \right| < 1 \ , \ z \in \boldsymbol{C}_1 \quad .$$

On the other hand, since $\bar{\boldsymbol{u}}_l(z)$ is the right-eigenvector of $\mathbf{Q}(z)$ associated with the eigenvalue $\lambda_l(z)$, we can write

$$|u_{il}(z)||\lambda_l(z)| = \left| \sum_{j=1}^{L} p_{ij} G_{ij}(z) u_{jl}(z) \right| \leq \sum_{j=1}^{L} p_{ij} \left| G_{ij}(z) \right| \cdot \left| u_{jl}(z) \right| \ , \ \forall z \in \boldsymbol{C}_1 \quad ,$$

which is valid for all $1 \leq i \leq L$. If, for each value of $z \in \boldsymbol{C}_1$, we now select i such that it corresponds to the $u_{il}(z)$ with the largest modulus, i.e., $|u_{il}(z)| \geq |u_{jl}(z)|$ for $j \neq i$, we then obtain

$$|\lambda_l(z)| \leq \sum_{j=1}^{L} p_{ij} \left| G_{ij}(z) \right|$$
$$< 1 \ , \ \forall z \in \boldsymbol{C}_1 \ .$$

□

It is noteworthy that this proof can be easily adapted to show that $|\lambda_i(z)| \leq 1$ for all z satisfying $|z| \leq 1$, implying that the eigenvalues of $\mathbf{Q}(z)$ are bounded for all values of z that belong to the complex unit disk $\{ z \in \mathbb{C} : |z| \leq 1 \}$.

Hence, for $H \to \infty$, convergence of the second term in right-hand side of the latter expression for $P_S(z,\bar{\boldsymbol{x}})$ is assured at the very least for those values of z that lie on the complex

unit circle $\{z\in\mathbb{C} : |z|=1 \wedge z\neq 1\}$, while the first term will vanish under these circumstances. As we will demonstrate before long, the contour $\boldsymbol{C}_1$ can be closed to include the value $z=1$. Working out the sum over h, we therefore obtain

$$\boxed{P_s(z,\bar{x}) = c(1-\rho)\sum_{\bar{m}} \frac{(z-1)(\Lambda(z)\mathbf{W}(z)\bar{x})^{\bar{m}}}{z^c - E_{\bar{m}}(z)}\Psi_{\bar{m}}(z)} \quad . \tag{III.18a}$$

$$\Psi_{\bar{m}}(z) \triangleq \sum_{\bar{l}} F_{\bar{l}\bar{m}}(z) \sum_{j=0}^{c-1} z^j p(j,\bar{l})$$

$$E_{\bar{m}}(z) \triangleq (\Lambda(z)\bar{I})^{\bar{m}} \equiv \prod_{i=1}^{L} \lambda_i(z)^{m_i} \quad .$$

This is the closed-form expression for the joint pgf of the system state that we sought for, and starting from this result, we will establish expressions for the generating functions of quantities such as the system and queue content, and the packet waiting time and delay. From this formula, we can extract the relevant performance measures concerning the buffer behaviour, such as mean, variance, and tail distribution of these quantities.

Elaborating on these considerations also allows us to establish an expression for $P_q(z,\bar{x})$, the joint pgf of the queue content at the beginning of an arbitrary slot, and the state of the packet arrival process during the preceding slot. Taking into account the relation (I.1b) between queue and system content, we can write

$$\begin{aligned} P_q(z,\bar{x}) &\triangleq \lim_{n\to\infty} \mathbf{E}\left[z^{q_{n+1}}\bar{x}^{\bar{a}_n}\right] = \mathbf{E}\left[z^{(s-c)^+}\bar{x}^{\bar{a}}\right] \\ &= z^{-c}\left\{\mathbf{E}\left[z^s\bar{x}^{\bar{a}}\right] + c(1-\rho)(z-1)\sum_{j=0}^{c-1} z^j \sum_{\bar{l}\in\Omega_{\bar{a}}} \bar{x}^{\bar{l}}\, p(j,\bar{l})\right\} . \end{aligned}$$

In addition, invoking $\mathbf{U}(z)\mathbf{W}(z)=\mathbf{I}$ (see (III.7a)) results in the following relation :

$$\bar{x}^{\bar{l}} = \prod_{i=1}^{L} x_i^{l_i} = \prod_{i=1}^{L}\left(\sum_{k=1}^{L} u_{ik}(z)\bar{w}_k(z)\bar{x}\right)^{l_i} = \sum_{\bar{m}\in\Omega_{\bar{a}}} F_{\bar{l}\bar{m}}(z)(\mathbf{W}(z)\bar{x})^{\bar{m}} \quad .$$

If we combine the latter two expressions with (III.18a), then via some elementary manipulations, it becomes obvious that $P_q(z,\bar{x})$ satisfies

$$\boxed{P_q(z,\bar{x}) = c(1-\rho)\sum_{\bar{m}\in\Omega_{\bar{a}}} \frac{(z-1)(\mathbf{W}(z)\bar{x})^{\bar{m}}}{z^c - E_{\bar{m}}(z)}\Psi_{\bar{m}}(z)} \tag{III.18b}$$

To conclude this section and check our calculations, let us set $z=1$, and verify that $P_S(1,\bar{\boldsymbol{x}})$ (or $P_q(1,\bar{\boldsymbol{x}})$), the steady-state joint pgf that describes the state of the underlying Markov chain during an arbitrary slot, indeed equals expression (III.10). Let us therefore define $\bar{\boldsymbol{N}}$ as the $L\times 1$ vector $[N\,0\,\cdots\,0]^T$; note that $\bar{\boldsymbol{N}}\in\Omega_{\bar{\boldsymbol{a}}}$ and is therefore one of the values taken by the summation index $\bar{\boldsymbol{m}}$ in the right-hand side of the expression for $P_S(z,\bar{\boldsymbol{x}})$. Then, since $\mathbf{Q}(1)$ corresponds to an irreducible and aperiodic Markov chain, the PF eigenvalue $\lambda_1(z)$ is the only eigenvalue satisfying $\lambda_1(1)=1$. Therefore, due to the occurrence of the factor $(z\text{-}1)$, the sum over $\bar{\boldsymbol{m}}$ in expression (III.18a) for $P_S(z,\bar{\boldsymbol{x}})$ will yield a nonzero term only if $\bar{\boldsymbol{m}}=\bar{\boldsymbol{N}}$. In addition, it can also be shown that $\lambda_1'(1)=E_i'(1)=c\rho/N$ (e.g. Appendix D), and we can therefore write

$$P_S(1,\bar{\boldsymbol{x}})=(\bar{\boldsymbol{w}}_1(1)\bar{\boldsymbol{x}})^N\sum_{\bar{\boldsymbol{l}}}F_{\bar{\boldsymbol{l}}\,\bar{\boldsymbol{N}}}(1)\sum_{j=0}^{c-1}p\left(j,\bar{\boldsymbol{l}}\right)\quad .$$

In view of expression (III.7d) for $\bar{\boldsymbol{w}}_1(1)$, the normalisation (III.16d) for the $p\left(j,\bar{\boldsymbol{l}}\right)$'s, and the identity $F_{\bar{\boldsymbol{l}}\,\bar{\boldsymbol{N}}}(1)=1$ that follows from combining (III.17a) with $\bar{\boldsymbol{u}}_1(1)=\bar{\boldsymbol{l}}$, we indeed obtain

$$P_S(1,\bar{\boldsymbol{x}})=\left(\bar{\boldsymbol{\pi}}^T\bar{\boldsymbol{x}}\right)^N=A(\bar{\boldsymbol{x}})\quad ,$$

as anticipated.

III.3.3 calculation of the boundary probabilities

The expression for $P_S(z,\bar{\boldsymbol{x}})$ (or $P_q(z,\bar{\boldsymbol{x}})$), in general, contains a total of $c\cdot|\Omega_{\bar{\boldsymbol{a}}}|$ boundary probabilities, represented by $p\left(j,\bar{\boldsymbol{l}}\right)$ defined in (III.16c), whose values are yet to be determined. Since $P_S(z,\bar{\boldsymbol{x}})$ is a polynomial of degree N for each of the variables $x_1,\cdots,x_N$, these arguments of $P_S(z,\bar{\boldsymbol{x}})$ can take any value within the complex plane. Consequently, similar to the i.i.d. case discussed in Section II-4, we can determine the boundary probabilities by expressing that $P_S(z,\bar{\boldsymbol{x}})$, as a (joint) probability generating function, is bounded whenever the complex variable z falls inside complex unit circle, i.e., $\{z\in\mathbb{C}:|z|\le 1\}$. In view of (III.18a), this indicates that the zeroes of the denominators in the right-hand side of these expression that fall inside the complex unit circle $\{z\in\mathbb{C}:|z|\le 1\}$, must necessarily be removable singular points of $P_S(z,\bar{\boldsymbol{x}})$, implying that such zeroes must also render the corresponding numerator equal to zero.

In Section III.3.2, we have shown that $|\lambda_i(z)| \leq 1$ for $\{ z \in \mathbb{C} : |z| \leq 1 \}$, implying that each of the functions $E_{\bar{m}}(z)$ satisfies $|E_{\bar{m}}(z)| \leq 1$ as well for such values of z, which points to the property that none of these functions has poles inside the complex unit circle. Unfortunately, as shown in Appendix D, (some of) the $\lambda_i(z)$'s may have *branch points* that fall inside the complex unit disk[(3)]. Hence, unlike the i.i.d. case, it is not possible to construct an irrefutable mathematical prove, based on Rouché's theorem, to show that each of the equations

$$z^c - E_{\bar{m}}(z) = 0 \;\; ; \;\; \bar{m} \in \Omega_{\bar{a}} \quad ,$$

necessarily has c solutions inside the complex unit disk. Nevertheless, although D-BMAP's in many shapes and forms have been studied and applied for a wide variety of traffic scenarios in the course of this research (e.g., [5], [7], [10], [12], [14], [31], [52], [82], [91], [114]), an example where this is not the case has not been encountered up to now, and, to the best of our knowledge, such an example, if it exists, has yet to be constructed. So for practical purposes, we can venture to state that the above equation has indeed c solutions inside the complex unit circle for each value of $\bar{m}$, provided that the equilibrium condition $\rho<1$ is fulfilled. We can even take this one step further with the premise that each of the equations

$$z = E_{\bar{m}}(z)^{1/c} \eta^i \;\; ; \;\; \bar{m} \in \Omega_{\bar{a}} \;\; , \; 0 \leq i \leq c-1 \quad , \tag{III.19a}$$

has exactly one solution inside the complex unit disk. Note that the $z=1$ is a zero of this equation for $\bar{m} = \bar{N}$ and $i=0$, since the PF eigenvalue $\lambda_1(z)$ satisfies $\lambda_1(1)=1$ (and is the only eigenvalue to do so), while the other zeroes will lie inside the complex unit circle $\{z : |z|<1\}$.

Let us therefore, for each value of $\bar{m}$, denote by $z_{i,\bar{m}}$, $0 \leq i \leq c$-1, the solution of equation (III.19a) inside the complex unit circle. By expressing that these quantities must be a zero of the respective numerators in expression (III.18a) for $P_S(z,\bar{x})$, we obtain a set of equations

$$\left(z_{i,\bar{m}} - 1\right) \sum_{\bar{l}} F_{\bar{l}\bar{m}}\left(z_{i,\bar{m}}\right) \sum_{j=0}^{c-1} z_{i,\bar{m}}^{j} \, p\left(j,\bar{l}\right) = 0 \;\; ; \;\; \bar{m} \in \Omega_{\bar{a}} \;\; , \; 0 \leq i \leq c-1 \quad , \tag{III.19b}$$

where the trivial equation for $z_{0,\bar{N}} = 1$ must be replaced by the normalisation condition (III.16c). This constitutes a set of $c \cdot |\Omega_{\bar{a}}|$ linear equations for the set of $c \cdot |\Omega_{\bar{a}}|$ unknown boundary probabilities that, in general, has a unique solution.

It should be pointed out however that some of the equations in (III.19b) may impose trivial relations on the $p\left(j,\bar{l}\right)$'s, which is strongly related to the property that some of these

---

(3) these branch points are removable singular points of $P_S(z,\bar{x})$; see Appendix D

boundary probabilities may perforce be equal to zero. This is best illustrated by considering a small example : assume that one of the states, say state S_i, of the underlying Markov chain is a *greedy state*, by which we mean that a source that visits this state during a slot will generate at least one packet in that slot. Remember that the entries of the probabilities $p\left(j,\overline{\boldsymbol{l}}\right)$ contain the event that the system content at the beginning of an arbitrary slot does not exceed j, combined with the state of the underlying Markov chain during the *foregoing* slot. Since packets that arrive during a slot are still part of the system content at the beginning of the next slot, we necessarily have that for such a scenario

$$p\left(j,\overline{\boldsymbol{l}}\right)=0 \Leftrightarrow l_i > j \quad ,$$

expressing that, if more than j sources are in the greedy state S_i during a slot, then the system content at the beginning of the next slot will necessarily exceed j as well. Hence, for each value of j and l_i, the number of nonzero boundary probabilities is equal to the number of different combinations by which N-l_i sources can visit the remaining L-1 non-greedy states, multiplied by c, and their total therefore equals, similar to (III.9b)

$$\sum_{j=0}^{c-1}\sum_{l_i=0}^{j}\binom{N+L-l_i-2}{N-l_i}=\sum_{l_i=0}^{c-1}(c-l_i)\binom{N+L-l_i-2}{N-l_i} \quad .$$

On the other hand, if at least one packet is generated in state S_i, then z will be a common factor of the i-the column of the pgm $\mathbf{Q}(z)$, which, in turn, implies that $z=0$ is a zero with multiplicity 1 of one of the L eigenvalues, say $\lambda_k(z)$ (hence, $\lambda_k(0)=0$ and $\lambda'_k(0)\neq 0$). Consequently, $z=0$ is a zero of the denominator $z^c - E_{\overline{\boldsymbol{m}}}(z)$ with multiplicity $\min\{c,m_k\}$. The nontrivial zeroes of the denominators are the ones that differ from 0, and their total equals

$$\sum_{m_k=0}^{N}(c-\min\{c,m_k\})\binom{N+L-m_k-2}{N-m_k}=\sum_{m_k=0}^{c-1}(c-m_k)\binom{N+L-m_k-2}{N-m_k} \quad .$$

We can therefore only conclude that, also in a traffic scenario with a greedy state, the number of non-trivial linear equations equals the number of non-zero boundary probabilities that need to be computed. A similar reasoning can be made if multiple greedy states exist, that each generates at least 1 (or more than 1) packet per slot. The final conclusion remains valid : in the end, we obtain a set of (non-trivial) linear equations for the same number of (non-zero) boundary probabilities that has a unique solution. The opposite of the general case with $c\cdot\left|\Omega_{\overline{\boldsymbol{a}}}\right|$ unknowns is the situation where the normalisation condition (III.16d) suffices for the calculation of just one single remaining unknown, which for instance occurs when, for c=1, all but one states are greedy states; e.g. [7], [137], [261], [265], [266] (note that if all states were greedy states, the equilibrium condition $\rho<1$ could never be met for $N\geq c$).

Nonetheless, it is clear that, depending on the values of N and L (and the number of non-greedy states), the number of non-zero boundary probabilities can become extremely large, in which case solving the set of linear equations (III.19b) becomes infeasible from a practical point-of-view. Also, when deriving numerical results for the performance measures considered further in this chapter, we evidently want to avoid all numerical calculations as much as possible. Therefore we also present an approximation for the boundary probabilities $p(j,\bar{\boldsymbol{l}})$ that reduces all numerical calculations to an absolute minimum. As before, we denote by the random variables e and s the number of packet arrivals during a tagged arbitrary slot and the system contents at the beginning of the following slot; also, $\bar{\boldsymbol{a}}$ represents the state of the Markovian arrival process during the tagged slot. We define the steady-state joint probability

$$q(j,\bar{\boldsymbol{l}}) \triangleq \Pr[e \le j, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}}] \quad .$$

Obviously, $s \le j$ implies that there have been at most j packet arrivals during the tagged slot, i.e. $s \le j \Rightarrow e \le j$. It is therefore clear that the inequality $q(j,\bar{\boldsymbol{l}}) \ge p(j,\bar{\boldsymbol{l}})$ holds for all $j \ge 0$, $\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}$. We will show by some numerical examples that approximating the boundary probabilities by

$$p(j,\bar{\boldsymbol{l}}) \cong C_a q(j,\bar{\boldsymbol{l}}) \ , \ C_a^{-1} \triangleq \sum_{j=0}^{c-1} \Pr[e \le j] = \sum_{j=0}^{c-1} (c-j) e(j) \quad , \tag{III.20a}$$

yields excellent approximations for the performance measures of interest. Observe that the factor C_a that appears in the right-hand side ensures that the normalisation condition (III.16c) is still valid when we substitute the $p(j,\bar{\boldsymbol{l}})$'s by their approximate values.

From expression (III.8), it follows that the steady-state joint pgf that describes the number of packet arrivals and the state of the Markov process in a slot can be expressed as

$$\sum_{i=0}^{\infty} z^i \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{l}}} \Pr\left[e=i, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}}\right] = \mathbf{E}\left[z^e \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}}\right] = \mathbf{E}\left[\left(\mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{a}}}\right] = \left(\bar{\boldsymbol{\pi}}^T \mathbf{Q}(z)\bar{\boldsymbol{x}}\right)^N \quad ,$$

and $\Pr\left[e=i, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}}\right]$ can, in principle, be computed by identifying the coefficient of z^j and $\bar{\boldsymbol{x}}^{\bar{\boldsymbol{l}}}$ in the right-hand side of this expression, which may be a complicated task. Fortunately, we do not need to know the value of each individual probability $q(j,\bar{\boldsymbol{l}})$, but rather in the specific combination that appears in the expression for $\Psi_{\bar{\boldsymbol{m}}}(z)$. First, combining definition (III.17a) and the above expression, we can derive that

$$\sum_{i=0}^{\infty} y^i \sum_{\bar{l},\bar{m}\in\Omega_{\bar{a}}} \bar{x}^{\bar{m}} F_{\bar{l}\bar{m}}(z) \Pr\left[e=i,\bar{a}=\bar{l}\right] = \sum_{i=0}^{\infty} y^i \sum_{\bar{l}\in\Omega_{\bar{a}}} \left(\mathbf{U}(z)\bar{x}\right)^{\bar{l}} \Pr\left[e=i,\bar{a}=\bar{l}\right]$$

$$= \left(\bar{\pi}^T \mathbf{Q}(y)\mathbf{U}(z)\bar{x}\right)^N \quad .$$

We then obtain an approximation $\Psi_{a,\bar{m}}(z)$ for $\Psi_{\bar{m}}(z)$, if we replace the $p\left(j,\bar{l}\right)$'s by their respective approximation $C_a \cdot q\left(j,\bar{l}\right)$:

$$\Psi_{\bar{m}}(z) \cong \Psi_{a,\bar{m}}(z) \triangleq C_a \sum_{j=0}^{c-1} z^j \sum_{\bar{l}\in\Omega_{\bar{a}}} F_{\bar{l}\bar{m}}(z) q\left(j,\bar{l}\right) \quad ,$$

which, in view of the penultimate relation, yields

$$\Psi_{a,\bar{m}}(z) = C_a \sum_{j=0}^{c-1} z^j \sum_{i=0}^{j} \theta_{i,\bar{m}}(z)$$

$$\theta_{i,\bar{m}}(z) \triangleq \frac{1}{i!}\frac{\partial^i}{\partial y^i}\left[\binom{N}{\bar{m}} \prod_{l=1}^{L} \left(\bar{\pi}^T \mathbf{Q}(y)\bar{u}_l(z)\right)^{m_l}\right]_{y=0} \quad ; \quad \binom{N}{\bar{m}} \triangleq \frac{N!}{m_1! \cdots m_L!} \quad . \tag{III.20b}$$

The advantage of adopting this approximation is two-fold : not only is there no need for the explicit numerical computation of neither the boundary probabilities $p\left(j,\bar{l}\right)$ nor the $q\left(j,\bar{l}\right)$'s that appear in their approximation (III.20a), it also prevents the calculation of the functions $F_{\bar{l}\bar{m}}(z)$ from equation (III.17a) that appear in expression (III.18a,b) for $P_s(z,\bar{x})$ and $P_q(z,\bar{x})$. The probabilities $\Pr[e \le j]$ in the definition of C_a, can be extracted from the expression for $E(z)$ given by (III.11). Mark that this approximation is particularly attractive in the single-server case c=1, since the calculation of derivatives (with respect to y for y=0) is then avoided, and we obtain a closed-form expression for $\Psi_{a,\bar{m}}(z)$ that is quite easy to handle. In case of an MMBP, these derivatives are relatively easy to calculate in closed-from for c>1.

As we will illustrate by means of some numerical examples later on, this approach to approximate the boundary probabilities can be further refined in the following way. To start, let us focus on the single-server case c=1. As indicated before, the condition s=0 that the buffer is empty during an arbitrary slot, is a (much) stronger one than the requirement e=0 of having no packet arrivals during the preceding slot – which is an arbitrary slot as well. On the other hand, s=0 implies that the server is 'idle', and we can imagine that the length of these idle periods can take quite a number of slots, especially when the correlation between the number of packet arrivals during successive slots is 'strong', which results in a 'bursty' packet arrival process : under these circumstances, we expect that the non-idle periods will be relatively long, implying that the idle periods will be relatively long as well. Hence, in such a

scenario, it is likely that an empty slot is preceded by *more than* one single slot during which no arrivals occur. Therefore, if we represent by e_M the number of packet arrivals during the M successive slots that precede an arbitrary slot, we will replace the condition s=0 by e_M=0, whereby the parameter M is to be determined later on by some (hopefully) simple criterion.

If we apply this scheme in a similar way as discussed above for M=1, this means that we approximate the boundary probabilities $p(0,\bar{\boldsymbol{l}})$ by $C_a \cdot q(0,\bar{\boldsymbol{l}})$, where the latter quantities are now defined as

$$q(0,\bar{\boldsymbol{l}}) \triangleq \Pr[\bar{\boldsymbol{a}}=\bar{\boldsymbol{l}}, e_M \le 0] \; ; \; C_a^{-1} \triangleq \Pr[e_M = 0] \quad ,$$

where e_M represents the total number of packet arrivals during M consecutive slots, and $\bar{\boldsymbol{a}}$ the state of the modulating Markov chain during the last of these M slots. Based on the previous results, it is then straightforward to show that the following relation holds :

$$\sum_{i=0}^{\infty} z^i \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{l}}} \Pr[e_M = i, \bar{\boldsymbol{a}} = \bar{\boldsymbol{l}}] = \left(\bar{\boldsymbol{\pi}}^T \mathbf{Q}(z)^M \bar{\boldsymbol{x}}\right)^N \quad ,$$

which, in turn, yields a closed-form approximate formula for the functions $\Psi_{\bar{\boldsymbol{m}}}(z)$

$$\Psi_{a,\bar{\boldsymbol{m}}}(z) = C_a \binom{N}{\bar{\boldsymbol{m}}} \prod_{l=1}^{L} \left(\bar{\boldsymbol{\pi}}^T \mathbf{Q}(0)^M \bar{\boldsymbol{u}}_l(z)\right)^{m_l} \; ; \; C_a^{-1} = \left(\bar{\boldsymbol{\pi}}^T \mathbf{Q}(0)^M \bar{\boldsymbol{1}}\right)^N \quad . \tag{III.21a}$$

where C_a^{-1} represents the probability that no packet is generated during a period of M consecutive slots. The quantities $\mathbf{Q}(0)^M$ that occur in this expression can be calculated through the decomposition of $\mathbf{Q}(z)$ into its eigenvalues and –vectors, as in (III.7b).

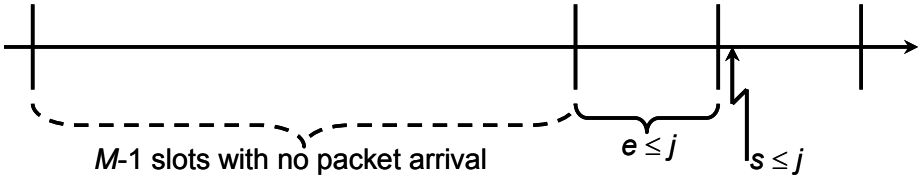


Figure III-2 : approximation of the boundary probabilities

Finally, by following an identical formal approach, this can be extended to the multi-server case if we replace the condition $s \le j$ at the beginning of an arbitrary slot, by the requirement that $e \le j$ during the previous slot, and no packets have arrived during the M-1 slots that precede the latter slot, which is represented by the condition $e_{M\text{-}1}$=0 in the remainder of this section; this is also clarified in Fig. III-2. This technique then leads to an approximation for the $\Psi_{\bar{\boldsymbol{m}}}(z)$'s that is similar to (III.20b), with the functions $\theta_{i,\bar{\boldsymbol{m}}}(z)$ and the normalisation constant C_a now satisfying

$$\theta_{i,\bar{\boldsymbol{m}}}(z) \triangleq \frac{1}{i!}\frac{\partial^i}{\partial y^i}\left[\binom{N}{\bar{\boldsymbol{m}}}\prod_{l=1}^{L}\left(\bar{\boldsymbol{\pi}}^T\mathbf{Q}(0)^{M-1}\mathbf{Q}(y)\bar{\boldsymbol{u}}_l(z)\right)^{m_l}\right]_{y=0}$$
$$C_a^{-1} = \sum_{i=0}^{c-1}(c-i)\frac{1}{i!}\frac{\partial^i}{\partial y^i}\left[\left(\bar{\boldsymbol{\pi}}^T\mathbf{Q}(0)^{M-1}\mathbf{Q}(y)\bar{\boldsymbol{I}}\right)^N\right]_{y=0} . \qquad \text{(III.21b)}$$

The relation between $s \le j$ at the start of an arbitrary slot on the one hand, and the requirement that the previous slot with $e \le j$ is preceded by M-1 slots with no arrivals, is less intuitively clear when $c>1$, which is why we expect that the approximations that we obtain in the multi-server case are less accurate. Nevertheless, the numerical examples that we consider in Section III.3.5.4 for the moments and in Section III.3.6.3 for the tail behaviour of the buffer content will reveal that the approximations that we thus obtain, are more than acceptable in terms of accuracy, on top of being computationally efficient as can be deduced from the above approximate formulas for $\Psi_{\bar{\boldsymbol{m}}}(z)$.

The final issue that needs to be resolved is to determine a good, if not optimal, choice for M. First, note that $\Pr[e \le j] > \Pr[s \le j]$, and that $\Pr[e \le j, e_{M-1}=0]$ becomes smaller as M increases. Consequently, we will determine M such that (the sum of) the probabilities of the events $s \le j$, $0 \le j \le c$-1, and their replacements $\{e \le j, e_{M\text{-}1}=0\}$, match as closely as possible

$$\sum_{j=0}^{c-1}\Pr[e \le j, e_{M-1}=0] \cong \sum_{j=0}^{c-1}\Pr[s \le j] = c(1-\rho) \quad ,$$

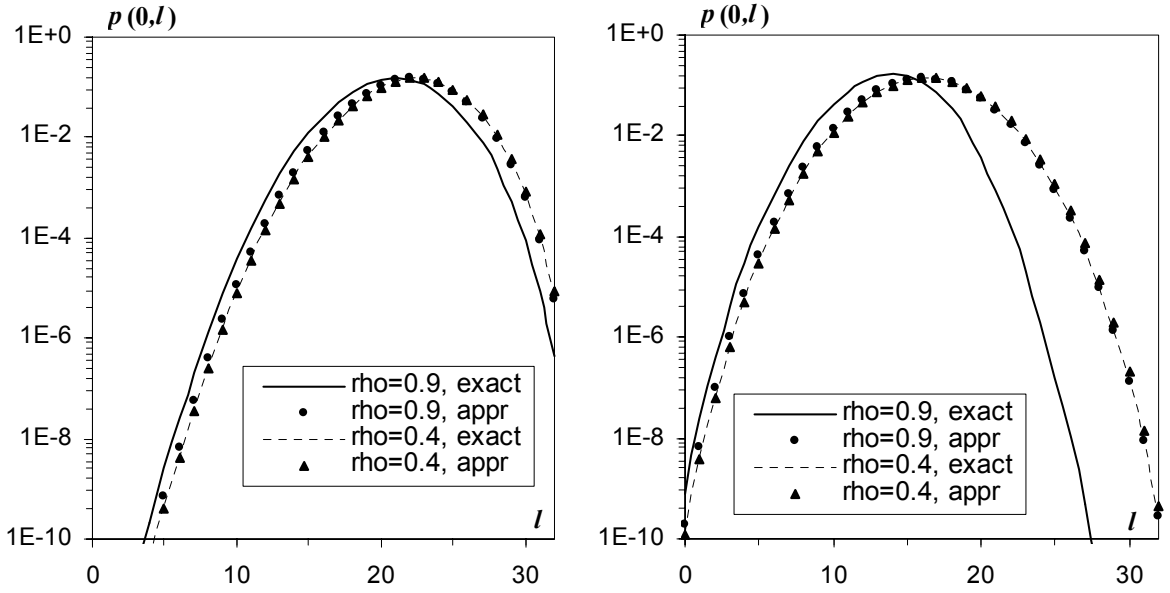
i.e., in view of expression (III.21b) for C_a, we will determine $M \ge 1$ to be the smallest integer for which

$$C_a^{-1} \le c(1-\rho) \quad . \qquad \text{(III.22)}$$

To illustrate the impact of this approach, let us compare the values of $p(j,\bar{\boldsymbol{I}})$ with their approximation for some specific values of the traffic parameters. We therefore describe each source by a 2-state ON/OFF process, with pgm

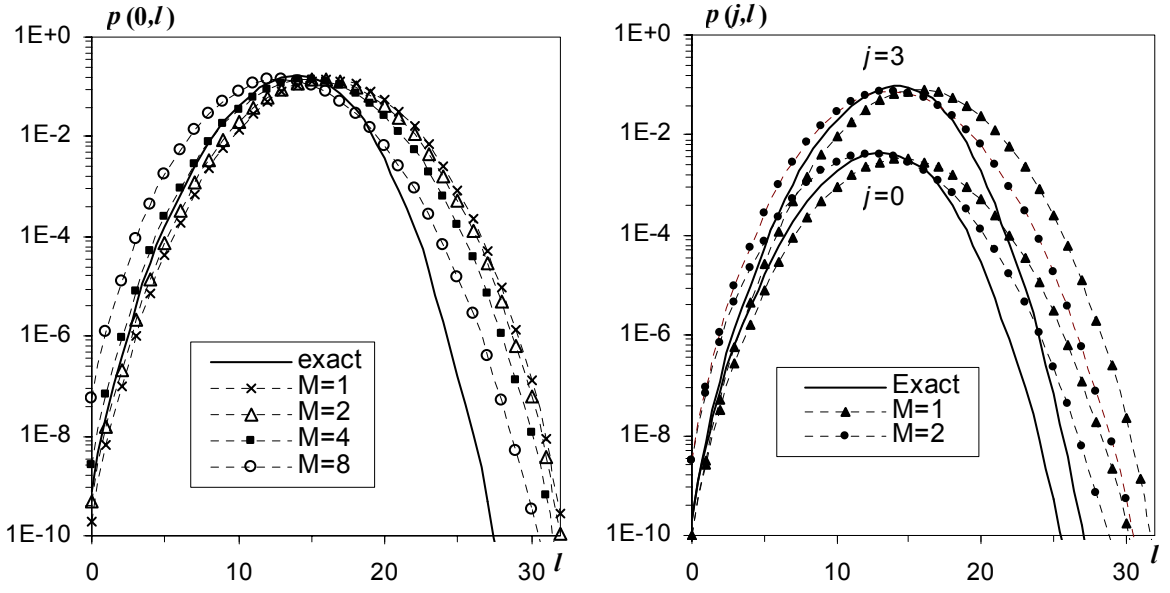
$$\mathbf{Q}(z) = \begin{bmatrix} (1-p_{12})G(z) & p_{12} \\ p_{21}G(z) & (1-p_{21}) \end{bmatrix} \; ; \; G(z) = 1-\gamma+\gamma z \quad , \qquad \text{(III.23)}$$

i.e., packets are generated according to a Bernoulli process with rate γ during each slot where state S_1 is visited, while no packets are generated when a source finds itself in state S_2 during a slot. Such a process is sometimes referred to as an *interrupted Bernoulli process*, or IBP. Evidently, a single source is characterised by a set of three parameters (γ, p_{12}, p_{21}), or,



(a) κ_b =2 (b) κ_b =20

Figure III-3 : $p(0,l)$ versus l; c=1, N=32, $\kappa_s=\kappa_b$



(a) c=1 (b) c=4

Figure III-4 : $p(j,l)$ versus l; $\kappa_s=\kappa_b$=20, ρ=0.9, N=32

equivalently, the set (ρ, κ_b, κ_s) with κ_c and κ_s as defined in Section III.2.4. How these two parameter sets are related will be discussed later on but is not so important at this point; we just need to remember that increasing values of κ_b (and, to a lesser extent, κ_s) correspond to increasingly capricious packet arrival streams, while κ_b=1 represents an uncorrelated – and even i.i.d. – arrival stream in case of an IBP. Also, in case of a 2×2 D-BMAP, we can represent the vector $\overline{\boldsymbol{l}}$ by its first component l, since $\overline{\boldsymbol{l}} = \begin{bmatrix} l & (N-l) \end{bmatrix}^T$.

In Figs. III-3a,b, for c=1, N=32, and κ_b=κ_s=2 (a) and 20 (b), we have plotted the boundary probabilities $p(0,l)$ and their approximation that follows from (III.20b) (i.e., with M=1), versus l, for ρ=0.4 and 0.9 respectively, as indicated. For low values of the burst factor, such as κ_s=2, there is little or no discernible difference between the values of *i*) the boundary probabilities for different values of the load ρ, and *ii*) the boundary probabilities and the proposed approximation. For high values of the burst factor, such as κ_s=20, then we still have an accurate approximation for low values of ρ, while the accuracy appears to deteriorate for high values of ρ – particularly when l is high as well. Nevertheless, observe that the approximation becomes less accurate when both the load ρ and the burst factor κ_b are high, which are exactly the circumstances under which the terms that contain the boundary probabilities in our expression for the mean and variance of the queue and/or system contents become less significant (e.g. the results of Section III.3.5); therefore we still expect this approximation to yield values that are sufficiently accurate under these traffic conditions.

In Fig. III-4a, considering the same (worst-case) parameter settings with ρ=0.9 and κ_s=20, we show the results for the approximation for the boundary probabilities that we obtain from (III.21b), for increasing values of M. While the accuracy deteriorates a bit for low values of l, it becomes better for high values of l. Similar results are shown in Fig. III-4b for the multi-server case with c=4, where we compare the values of $p(j,l)$ with their respective approximation, for j=0,3 and M=1,2. These results indicate that, although there may still be a discrepancy between the approximate and exact data, the overall accuracy of the proposed approximation procedure does improve for increasing M, as we hoped for. This will be confirmed when we investigate the impact of these approximations on the value of the performance measures such as the mean, variance, and tail distribution of the queue and system content, which will be discussed later on when expressions for these quantities will be derived.

III.3.4 establishing expressions for $S(z)$ and $Q(z)$

In view of the definition of $P_S(z,\vec{x})$, an expression for $S(z)$, the steady-state pgf of the system content at the beginning of an arbitrary slot, simply follows from the identity $S(z) \equiv P_S(z,\vec{1})$. If we then rely on the property that $\mathbf{W}(z)\vec{1} = \vec{1}$ – this is one of the key points where our preference for the 'special' solution for the eigenvectors in (III.7a,b,c) pays off – then we easily obtain from (III.18a)

$$\boxed{S(z) = c(1-\rho)\sum_{\vec{m}} \frac{(z-1)E_{\vec{m}}(z)}{z^c - E_{\vec{m}}(z)}\Psi_{\vec{m}}(z)} \qquad \text{(III.24)}$$

An expression for $Q(z)$, on the other hand, can be similarly established from $Q(z) \equiv P_q(z, \bar{\boldsymbol{l}})$, which eventually produces the following formula :

$$\boxed{Q(z) = \sum_{\bar{\boldsymbol{m}}} \frac{c(1-\rho)(z-1)}{z^c - E_{\bar{\boldsymbol{m}}}(z)} \Psi_{\bar{\boldsymbol{m}}}(z)} \quad . \tag{III.25}$$

It is interesting to note the formal resemblance between these formulae for $S(z)$ and $Q(z)$, and expression (II.10) and (II.13) for $S(z)$ and $Q(z)$ respectively in case of an i.i.d. packet arrival process. This likeness will be fully exploited when we derive expressions for the mean, variance, and tail distribution of the queue and system content in the subsequent sections.

In order to check our derivations yet again, let us consider the parameter set of expression (III.13a), for which the arrival pattern is an i.i.d. process, and verify that the above expressions for $S(z)$ and $Q(z)$ are indeed reduced to their respective i.i.d. counterpart, for which expressions were established in Chapter II. As already pointed out in Section III.2.3, the pgm $\mathbf{Q}(z)$ of (III.13a) generates an i.i.d. arrival pattern, since the set of equations (III.12a) are indeed fulfilled. Furthermore, the solution of the characteristic equation of this pgm leads to $\lambda_1(z) = E_i(z)$, as expected, and $\lambda_2(z) \equiv 0$, while the right eigenvectors are given by (see Appendix D)

$$\mathbf{U}(z) \equiv \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \ \forall z \quad .$$

Therefore, the only term in the sum over $\bar{\boldsymbol{m}}$ in the expressions for $S(z)$ (and $Q(z)$) that is nonzero is for $\bar{\boldsymbol{m}} = \bar{\boldsymbol{N}}$. In addition, since $F_{\bar{\boldsymbol{l}}\,\bar{\boldsymbol{N}}}(z) \equiv 1$ under these circumstances, we find

$$S(z) = c(1-\rho)\frac{(z-1)E_i(z)^N}{z^c - E_i(z)^N}\Psi_{\bar{\boldsymbol{N}}}(z) \ ; \ \Psi_{\bar{\boldsymbol{N}}}(z) = \sum_{\bar{\boldsymbol{l}}} \sum_{j=0}^{c-1} z^j p(j, \bar{\boldsymbol{l}}) = \frac{1}{c(1-\rho)} \sum_{j=0}^{c-1} z^j \Pr[s \le j] \quad ,$$

which indeed corresponds to the steady-state pgf of the system content in case of i.i.d. arrivals; see expressions (II.10) and (II.23). A similar result evidently also holds for $Q(z)$.

III.3.5 the moments of the buffer content

III.3.5.1 closed-form expressions for the mean and variance

From the expressions that were established in the previous section for the steady-state pgfs of the queue and system content, we can now derive closed-form expressions for the performance indices of these quantities. Let us commence by calculating μ_q, by applying the

$\mathcal{M}[\cdot]$-operator on formula (III.25) for $Q(z)$. First, observe that, in view of definition (III.17a), and the solution (III.7d) for $\mathbf{U}(z)$ in case of $z\to 1$, the $F_{\bar{l}\bar{m}}(z)$-functions at $z=1$ become

$$F_{\bar{l}\bar{m}}(1)=\begin{cases}1\ , & \bar{m}=\bar{N}\\ 0\ , & \forall\,\bar{m}\in\Omega_{\bar{a}}\setminus\bar{N}\end{cases}\ ;\ \forall\,\bar{l}\in\Omega_{\bar{a}}\quad , \tag{III.26a}$$

while we also find that

$$F_{\bar{l}\bar{N}}(z)=\prod_{i=1}^{L}u_{i1}(z)^{l_i}\quad , \tag{III.26b}$$

an expression that we will rely upon in the subsequent derivations. The identities (III.26a), among others, imply that

$$\Psi_{\bar{m}}(1)=\begin{cases}1\ , & \bar{m}=\bar{N}\\ 0\ , & \forall\,\bar{m}\in\Omega_{\bar{a}}\setminus\bar{N}\end{cases}\quad . \tag{III.26c}$$

Therefore, in view of this relation and the factor (z-1) that accompanies each of these terms in the sum over $\bar{m}$ that appears in the right-hand side of expression (III.25), we can state

$$\frac{\partial}{\partial z}\frac{c(1-\rho)(z-1)}{z^c-E_{\bar{m}}(z)}\Psi_{\bar{m}}(z)\bigg|_{z=1}=\mathcal{M}\left[\frac{c(1-\rho)(z-1)}{z^c-E_{\bar{m}}(z)}\Psi_{\bar{m}}(z)\right]=0\ ;\ \forall\,\bar{m}\in\Omega_{\bar{a}}\setminus\bar{N}\quad , \tag{III.26d}$$

and the only term that yields a nonzero contribution in the calculation of μ_q is the one for $\bar{m}=\bar{N}$. Consequently, due to $\lambda_1(1)=1$ and $N\lambda_1'(1)=\mu_e$, definition (III.15b) for κ_b, and the similarity between (III.25) and (II.13) for $Q(z)$, we immediately obtain (e.g. (II.19))

$$\mu_q=\frac{\kappa_b\sigma_e^2}{2(c-\mu_e)}-\frac{\mu_e}{2}+\mathcal{M}\left[\Psi_{\bar{N}}\right]-\frac{c-1}{2}\quad , \tag{III.27a}$$

where

$$\mathcal{M}\left[\Psi_{\bar{N}}\right]=\sum_{j=0}^{c-1}\sum_{\bar{l}}\mathcal{M}\left[z^jF_{\bar{l}\bar{N}}(z)\right]p\left(j,\bar{l}\right)=\sum_{j=0}^{c-1}\sum_{\bar{l}}\left(j+\sum_{i=1}^{L}l_iu_{i1}'(1)\right)p\left(j,\bar{l}\right)\quad . \tag{III.27b}$$

For the calculation of expressions for the derivatives of the PF eigenvalue and right-eigenvector with respect to z at z=1, we refer to the results and remarks listed in Appendix D.

An expression for the variance of the queue content can be obtained in a similar manner. First, observe that the general computation rules (II.15a) that apply for $\mathcal{V}[\cdot]$, combined with the identities (III.26c), yield

$$\sigma_q^2 = \boldsymbol{\mathcal{V}}\left[\frac{c(1-\rho)(z-1)}{z^c - E_{\bar{\boldsymbol{N}}}(z)}\Psi_{\bar{\boldsymbol{N}}}(z)\right] + \sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{\boldsymbol{N}}} \boldsymbol{\mathcal{V}}\left[\frac{c(1-\rho)(z-1)}{z^c - E_{\bar{\boldsymbol{m}}}(z)}\Psi_{\bar{\boldsymbol{m}}}(z)\right]$$
$$= \boldsymbol{\mathcal{V}}\left[\frac{c(1-\rho)(z-1)}{z^c - E_{\bar{\boldsymbol{N}}}(z)}\Psi_{\bar{\boldsymbol{N}}}(z)\right] + 2\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{\boldsymbol{N}}} \frac{c(1-\rho)}{1-E_{\bar{\boldsymbol{m}}}(1)}\boldsymbol{\mathcal{M}}[\Psi_{\bar{\boldsymbol{m}}}] \ .$$

Due to the similarity between the computation of the first term in the right-hand side of this expression, and the calculation of σ_q^2 in the case of i.i.d. arrivals (e.g., (II.21)), we may write

$$\sigma_q^2 = \frac{\kappa_s \mu_{3,e}}{3(c-\mu_e)} + \left(\frac{\kappa_b \sigma_e^2}{2(c-\mu_e)}\right)^2 - \frac{\kappa_b\sigma_e^2}{2} + \frac{1-(c-\mu_e)^2}{12} + \boldsymbol{\mathcal{V}}[\Psi_{\bar{\boldsymbol{N}}}] + 2\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{\boldsymbol{N}}} \frac{c(1-\rho)}{1-E_{\bar{\boldsymbol{m}}}(1)}\boldsymbol{\mathcal{M}}[\Psi_{\bar{\boldsymbol{m}}}] \ . \quad \text{(III.28a)}$$

Once more invoking the computational rules (II.15a), then the term $\boldsymbol{\mathcal{V}}[\Psi_{\bar{\boldsymbol{N}}}]$ in (III.28a) can be converted into

$$\begin{aligned}\boldsymbol{\mathcal{V}}[\Psi_{\bar{\boldsymbol{N}}}] &= \sum_{j=0}^{c-1}\sum_{\bar{\boldsymbol{l}}}\left(\boldsymbol{\mathcal{V}}\left[z^j F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{N}}}(z)\right] + \boldsymbol{\mathcal{M}}\left[z^j F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{N}}}(z)\right]^2\right)p(j,\bar{\boldsymbol{l}}) - \boldsymbol{\mathcal{M}}[\Psi_{\bar{\boldsymbol{N}}}]^2 \\ &= \sum_{j=0}^{c-1}\sum_{\bar{\boldsymbol{l}}}\left\{\sum_{i=1}^{L} l_i\left(u_{i1}''(1) + u_{i1}'(1) - u_{i1}'(1)^2\right) + \left(j + \sum_{i=1}^{L} l_i u_{i1}'(1)\right)^2\right\}p(j,\bar{\boldsymbol{l}}) - \boldsymbol{\mathcal{M}}[\Psi_{\bar{\boldsymbol{N}}}]^2 .\end{aligned} \quad \text{(III.28b)}$$

We need some additional derivations to bring the last term in the right-hand side of (III.28a) in a form that avoids the explicit calculation of the derivatives of the $F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)$–functions at $z=1$. First, note that this term can be written as

$$\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{\boldsymbol{N}}} \frac{c(1-\rho)}{1-E_{\bar{\boldsymbol{m}}}(1)}\boldsymbol{\mathcal{M}}[\Psi_{\bar{\boldsymbol{m}}}] = c(1-\rho)\sum_{j=0}^{c-1}\sum_{\bar{\boldsymbol{l}}\in\Omega_{\bar{\boldsymbol{a}}}} \Delta_{\bar{\boldsymbol{l}}}\, p(j,\bar{\boldsymbol{l}})$$
$$\Delta_{\bar{\boldsymbol{l}}} \triangleq \frac{\partial}{\partial z}\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{\boldsymbol{N}}} \frac{F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)}{1-E_{\bar{\boldsymbol{m}}}(1)}\bigg|_{z=1} \ .$$

Due to definitions (III.18a) for $E_{\bar{\boldsymbol{m}}}(z)$ and (III.17a) for $F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)$, the coefficients $\Delta_{\bar{\boldsymbol{l}}}$ can be converted into

$$\Delta_{\bar{\boldsymbol{l}}} = \frac{\partial}{\partial z}\sum_{h=0}^{\infty}\left(\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)\left(\boldsymbol{\Lambda}(1)^h\bar{\boldsymbol{l}}\right)^{\bar{\boldsymbol{m}}} - F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{N}}}(z)\right)\bigg|_{z=1}$$
$$= \frac{\partial}{\partial z}\sum_{h=0}^{\infty}\left(\prod_{i=1}^{L}\left(\sum_{k=1}^{L} u_{ik}(z)\lambda_k(1)^h\right)^{l_i} - \prod_{i=1}^{L} u_{i1}(z)^{l_i}\right)\bigg|_{z=1} ,$$

where we in particular have made use of the property that $|E_{\bar{\boldsymbol{m}}}(1)|<1$ if $\bar{\boldsymbol{m}}\neq\bar{\boldsymbol{N}}$ to ensure convergence of the sum over h. Working out the derivative with respect to z, this becomes

$$\Delta_{\bar{\boldsymbol{l}}} = \sum_{h=0}^{\infty} \sum_{i=1}^{L} l_i \left(\sum_{k=1}^{L} u'_{ik}(1) \lambda_k(1)^h - u'_{i1}(1) \right) = \sum_{h=0}^{\infty} \sum_{i=1}^{L} l_i \left(\sum_{k=2}^{L} u'_{ik}(1) \lambda_k(1)^h \right) \quad .$$

Since the non-PF eigenvalues satisfy $|\lambda_k(1)|<1$ for $2 \leq k \leq L$, the sum over h in this expression converges, and we thus finally obtain

$$\sum_{\bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}} \setminus \bar{\boldsymbol{N}}} \frac{c(1-\rho)}{1-E_{\bar{\boldsymbol{m}}}(1)} \mathcal{M}[\Psi_{\bar{\boldsymbol{m}}}] = c(1-\rho) \sum_{j=0}^{c-1} \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} \left(\sum_{i=1}^{L} l_i \sum_{k=2}^{L} \frac{u'_{ik}(1)}{1-\lambda_k(1)} \right) p(j, \bar{\boldsymbol{l}}) \quad . \qquad \text{(III.28c)}$$

The combination of (III.28a–c) thus enables the numerical computation of the variance of the queue content, primarily in terms of the PF-eigenvalue and right-eigenvector – in particular their derivatives at $z=1$ – although the sum in (III.28c) requires some information concerning the non-PF-eigenvalues and –vectors as well. These results exemplify the relevance and usefulness of our definition of the burst factor κ_b and κ_s in Section III.2.4. Apparently, if we are able to compute / measure / estimate the first three moments of the total number of packet arrivals over a long time interval (divided by the length of the interval), then these quantities determine the moments of the queue and system content in a similar manner, as the moments of the number of packet arrivals per slot in case of an i.i.d. packet arrival process.

Finally, the above procedure can now be repeated as a whole to calculate the mean and variance of the system content as well. It can then be shown that the following relations hold between the first two moments of the system and queue content :

$$\begin{aligned} \mu_s &= \mu_q + \mu_e \\ \sigma_s^2 &= \sigma_q^2 + \kappa_b \sigma_e^2 + 2c(1-\rho) \sum_{j=0}^{c-1} \sum_{\bar{\boldsymbol{l}} \in \Omega_{\bar{\boldsymbol{a}}}} \left(\sum_{i=1}^{L} l_i u'_{i1}(1) \right) p(j, \bar{\boldsymbol{l}}) \end{aligned} \quad . \qquad \text{(III.29)}$$

The relation between the mean of the queue and system content immediately follows from the relation (II.11b) established in the previous chapter which relates the queue and system content at consecutive slot boundaries, which naturally remains valid in case of a non-independent packet arrival process during consecutive slots.

III.3.5.2 an approximation for the moments of the buffer content

If we focus on $\mathcal{M}[\Psi_{\bar{\boldsymbol{N}}}]$ given by expression (III.27b), then we must come to the conclusion that it is not possible to impose strict bounds on the value of this quantity, due to the presence of the derivatives of the PF-eigenvalue at $z=1$ in this expression. Indeed, notwithstanding equations such as (see the results for $L=2$ in Appendix D)

` $\bar{\boldsymbol{\pi}}^T \bar{\boldsymbol{u}}'_k(1) = 0 \quad ; \quad 1 \leq k \leq L \quad ,$

the individual values of $u'_{k1}(1)$ can nonetheless take any particular value, depending on the system parameters that describe the packet arrival process. Nonetheless, the extensive set of numerical examples that will be considered later on will demonstrate that if we replace $\Psi_{\bar{\boldsymbol{m}}}(z)$ by $\Psi_{a,\bar{\boldsymbol{m}}}(z)$ given by (III.21a), then the outcome is an approximation for the performance measures that is fairly accurate and very easy to handle, in view of the closed-form nature of the expressions that we obtain by adopting the procedures proposed in Section-III.3.3.

Let us illustrate this for the single-server case $c=1$. With the use of the formulae presented in Section III.3.3, we then obtain after some elementary calculations

$$\begin{aligned} \mathcal{M}\left[\Psi_{\bar{N}}\right] &\cong \mathcal{M}\left[\Psi_{a,\bar{N}}\right] = C_a^{1/N} N\left(\bar{\pi}^T \mathbf{Q}(0)^M \bar{u}'_1(1)\right) \\ \mathcal{V}\left[\Psi_{\bar{N}}\right] &\cong \mathcal{V}\left[\Psi_{a,\bar{N}}\right] = C_a^{1/N} N\left(\bar{\pi}^T \mathbf{Q}(0)^M \bar{u}''_1(1)\right) + \mathcal{M}\left[\Psi_{a,\bar{N}}\right]\left(1-\mathcal{M}\left[\Psi_{a,\bar{N}}\right]/N\right) \end{aligned} \quad , \qquad \text{(III.30a)}$$

where $C_a^{-1/N}$ is easily calculated from (III.21a). In addition, observe that the first derivative with respect to z of $\Psi_{a,\bar{\boldsymbol{m}}}(z)$ for $z=1$ yields a nonzero result only if either $m_1=N$ or $m_1=N-1$ (with m_1 the first component of $\bar{\boldsymbol{m}}$). Hence, if we replace $\Psi_{\bar{\boldsymbol{m}}}(z)$ by $\Psi_{a,\bar{\boldsymbol{m}}}(z)$ in (III.28c), we find the following approximation for this expression :

$$\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{N}} \frac{(1-\rho)}{1-E_{\bar{\boldsymbol{m}}}(1)} \mathcal{M}\left[\Psi_{\bar{\boldsymbol{m}}}\right] \cong N C_a^{1/N} (1-\rho) \sum_{k=2}^{L} \frac{\bar{\pi}^T \mathbf{Q}(0)^M \bar{u}'_k(1)}{1-\lambda_k(1)} \quad . \qquad \text{(III.30b)}$$

The formulae that we obtain for these quantities are somewhat more involved in the multi-server case $c>1$, but lend themselves equally well for implementation in a numerical procedure.

III.3.5.3 heavy-load approximations

From the expressions for the mean and variance of the queue or system content, we can extract a heavy-load approximation in a entirely analogous way as was demonstrated in Section II.3.2.3 of the previous chapter. As ρ comes closer to 1, the dominant terms in these expressions will be those that have $\rho=1$ as pole (with multiplicity 2 in case of the variance), and we obtain

$$\begin{aligned} \mu_{q,\rho\to 1} &= \mu_{s,\rho\to 1} = \frac{\kappa_b \sigma_e^2}{2c(1-\rho)} \\ \sigma^2_{q,\rho\to 1} &= \sigma^2_{s,\rho\to 1} = \left(\frac{\kappa_b \sigma_e^2}{2c(1-\rho)}\right)^2 \end{aligned} \quad ,$$

Apparently, for fixed values of the load ρ (close to 1), the heavy-load behaviour of both the mean and variance of the queue and buffer content is determined by the variance of the packet arrival process *measured over a long time period*, which can be regarded upon as a measure for the variability, or burstiness, that is induced in the arrival process by the D-BMAP.

From these results, we can already conclude that the burst factor κ_b will play a similar role as the clumping factor κ_c of the previous chapter, since, for high enough values of the load ρ, both the mean and standard deviation of the system and queue content are proportional to this parameter. In addition, we also observe that the coefficient of variation of the queue and buffer content once again tends to 1 as the load ρ becomes close to 1

$$C_{V,\rho\to 1} = \frac{\sigma_{q,\rho\to 1}}{\mu_{q,\rho\to 1}} = \frac{\sigma_{s,\rho\to 1}}{\mu_{s,\rho\to 1}} = 1 \quad ,$$

similar to the i.i.d.-arrivals case treated in Section III.3.5.3.

III.3.5.4 numerical examples

Let us, for the time being, confine ourselves to an IBP arrival pattern, defined by the pgm (III.23). The stationary vector that corresponds to such a process, which is the solution of the set of equations defined by (III.4a), satisfies

$$\pi_1 = \frac{p_{21}}{p_{12}+p_{21}} \; ; \; \pi_2 = \frac{p_{12}}{p_{12}+p_{21}} \quad ,$$

and the time during which a source resides in the states S_1 and S_2 is geometrically distributed with mean $1/p_{12}$ and $1/p_{21}$ respectively. The source behaviour can be characterised by the set of three parameters (p_{12}, p_{21}, γ), or, equivalently, the set of parameters (ρ, κ_b, κ_s). From the results that are summarised in Appendix D, we can deduce that the latter set of parameters can be expressed in terms of the former one as

$$\frac{c\rho}{N} = \frac{p_{21}}{p_{12}+p_{21}}\gamma = \pi_1\gamma \quad , \tag{III.31a}$$

in addition to

$$\begin{aligned} \kappa_b &= 1 + 2\frac{\pi_1\gamma}{1-\pi_1\gamma}\frac{p_{12}}{p_{21}}\left(\frac{1-p_{12}-p_{21}}{p_{12}+p_{21}}\right) \\ \kappa_s &= 1 + 3\frac{\kappa_b - 1}{1-2\pi_1\gamma}\left((1-\pi_1\gamma)\frac{\kappa_b+1}{2} - \frac{\pi_1\gamma}{p_{12}+p_{21}}\right) \end{aligned} \quad . \tag{III.31a}$$

Inversely, defining a and b as

$$a \triangleq \frac{(\kappa_b - 1)(1-\pi_1\gamma)}{2\pi_1\gamma} \; ; \; b \triangleq \frac{(\kappa_b + 1)(1-\pi_1\gamma)}{2\pi_1\gamma} - \frac{(\kappa_s - 1)(1-2\pi_1\gamma)}{3\pi_1\gamma(\kappa_b - 1)} \quad ,$$

we find that

$$p_{12} = \frac{a}{b(a+b-1)} \; ; \; p_{21} = \frac{b-1}{b(a+b-1)} \; ; \; \gamma = \frac{c\rho}{N}\frac{p_{12}+p_{21}}{p_{21}} \quad . \tag{III.31b}$$

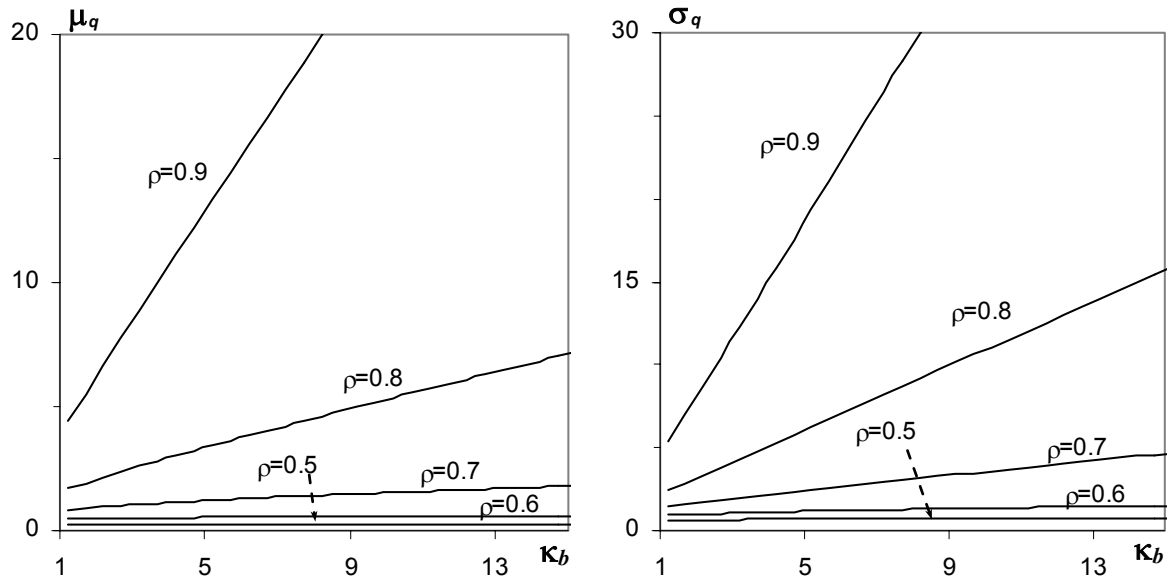
From expressions (III.31a), we derive that the only non-trivial parameter setting[4] for which $\kappa_b=\kappa_s=1$ occurs when $p_{12}+p_{21}=1$, and it is not difficult to check that such a solution fulfils the set of equations (III.12a), and thus represents an i.i.d. arrival process. On the other hand, from condition (III.14), one readily deduces that the arrival process will be uncorrelated – but not necessarily independent – if the relation $(1-p_{12})\gamma=c\rho/N$ holds, and a simple calculation reveals that this condition is indeed satisfied if $p_{12}+p_{21}=1$, which confirms that an independent arrival process is an uncorrelated one as well, as anticipated.

In the numerical examples that follow hereafter, we focus on the queue content, since the results for the system content look quite similar.

Figs. III-5a,b depict the mean (a) and stdv (b) of the queue content versus the burst factor κ_b, for $\kappa_s=1.5$, $N=16$ and $c=1$, and various values of the load ρ as indicated. A conspicuous property of these two performance indices is their quasi-linear behaviour as a function of the value of the burst factor. Whereas there is no perceptible dependence for low values of the load ρ (say lower than 0.6), this dependence becomes more pronounced as ρ increases, and is very strong for high values of ρ. This the region where the mean and stdv of the queue content is for the largest part determined by the heavy-load behaviour mentioned in Section III.3.5.3, where the linear dependence of μ_q and σ_q on the value of κ_b becomes obvious. In Figs. III-6a,b we have plotted analogous results for the multi-server case $c=4$, with $N=64$, thereby illustrating that a close-to-linear dependence of μ_q and σ_q on the value of κ_b exists in the multi-server case as well. Whereas the observed behaviour is quite similar to the single-server scenario for low values of the load ρ, a much weaker dependence of μ_q and σ_q on the value of the burst factor exists for higher values of ρ; this appears to be particular true for intermediate values of ρ (e.g., the results for $\rho=0.7$ and 0.8). Hence, a multi-server queuing system appears to be better able to handle capricious arrival patterns, compared to the single-server case. Nevertheless, when the load is (too) high, large buffers will have to be provided in any case to absorb the bursts of packets that enter the system.

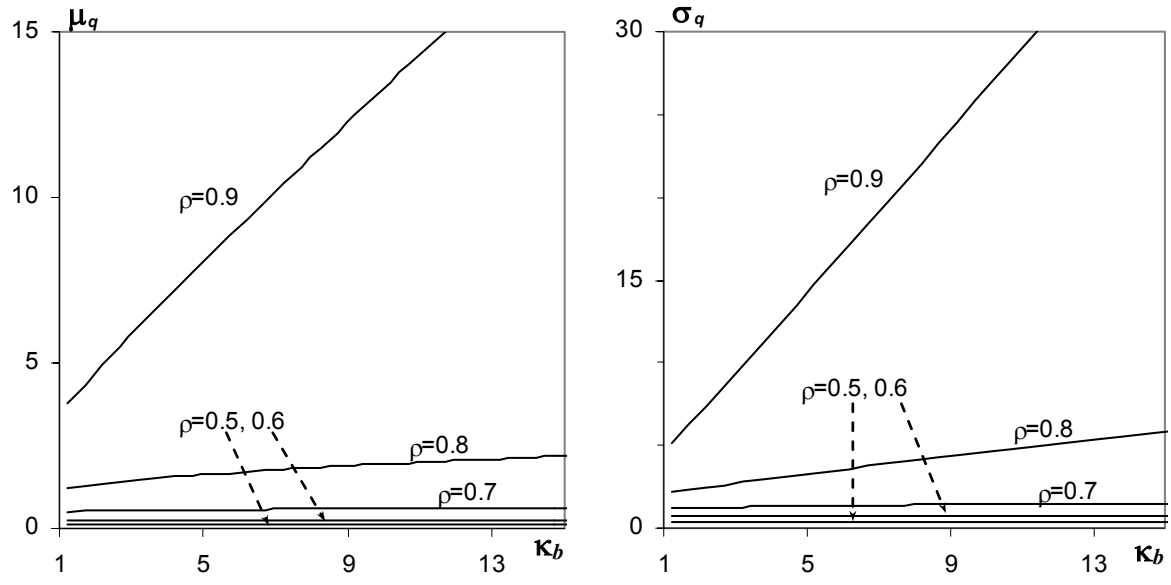
These observation are confirmed by the data presented in Figs. III-7a,b, where we show the mean (a) and stdv (b) of the queue content, relative to the value of these quantities in case the

[4] a parameter value such as $p_{12}=0$ yields a non-recurrent 2-state Markov chain, a scenario that we wish to exclude from our analyses



(a) mean (b) stdv

Figure III-5 : moments of the queue content versus κ_b; c=1, N=16, κ_s=1.5



(a) mean (b) stdv

Figure III-6 : moments of the queue content versus κ_b; c=4, N=64, κ_s=1.5

packet arrival process were an i.i.d. process with identical values of (ρ, c, N), versus the load ρ, for c=4 and N=64, and values of the burst factor κ_b=2,3,4,5. As explained before, the mean and stdv of the queue and system contents in case of i.i.d. arrivals are the results for these quantities that correspond to κ_b=1 (=κ_s). For low values of ρ, there is no discernable impact of the value of the burst factor, while for $\rho \to 1$, these ratios converge to κ_b. Hence, a safe – albeit rather coarse if ρ is not too high – approximation for the (first two) moments of the queue content could consist of multiplying the values for these moments that are

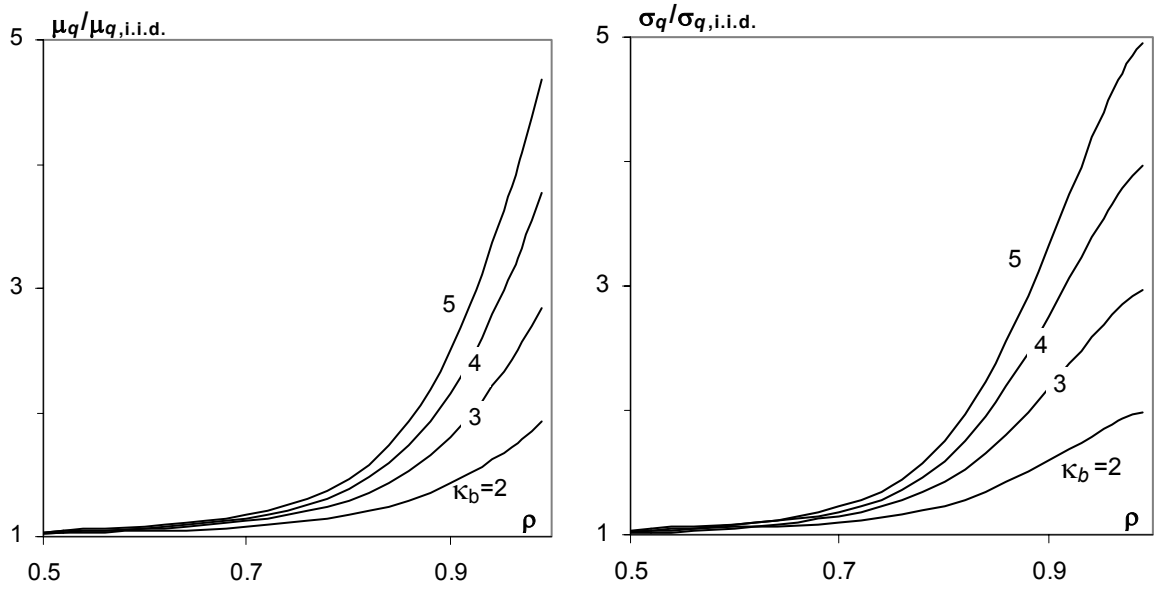


Figure III-7 : moments of the queue content, relative to the i.i.d. case, versus ρ; c=4, N=64, $\kappa_s=\kappa_b$

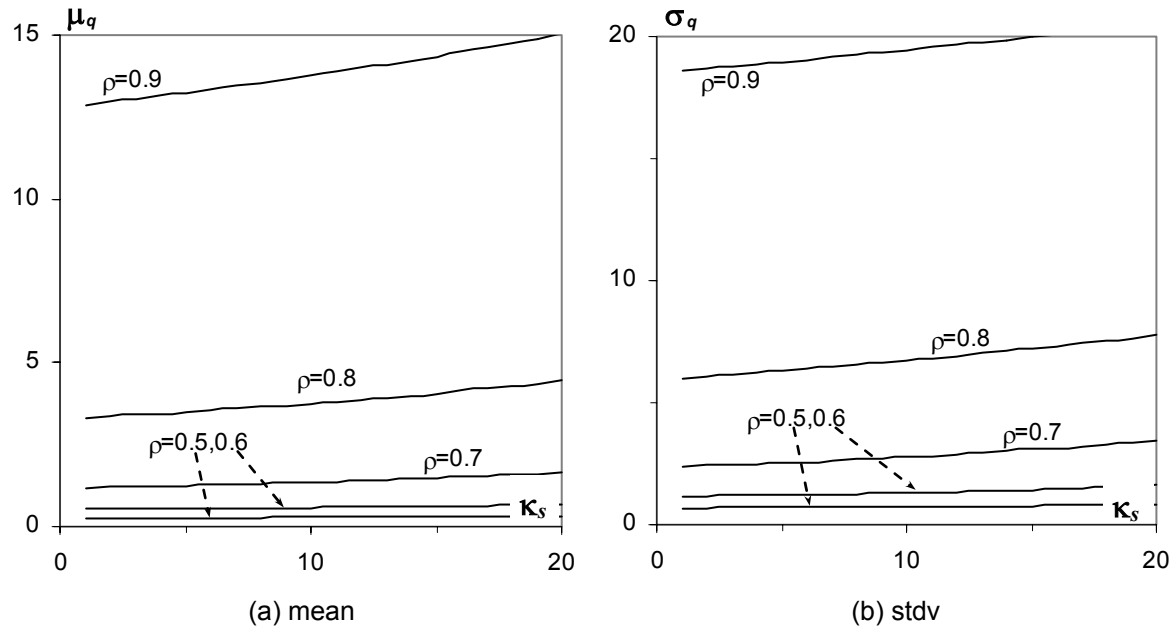
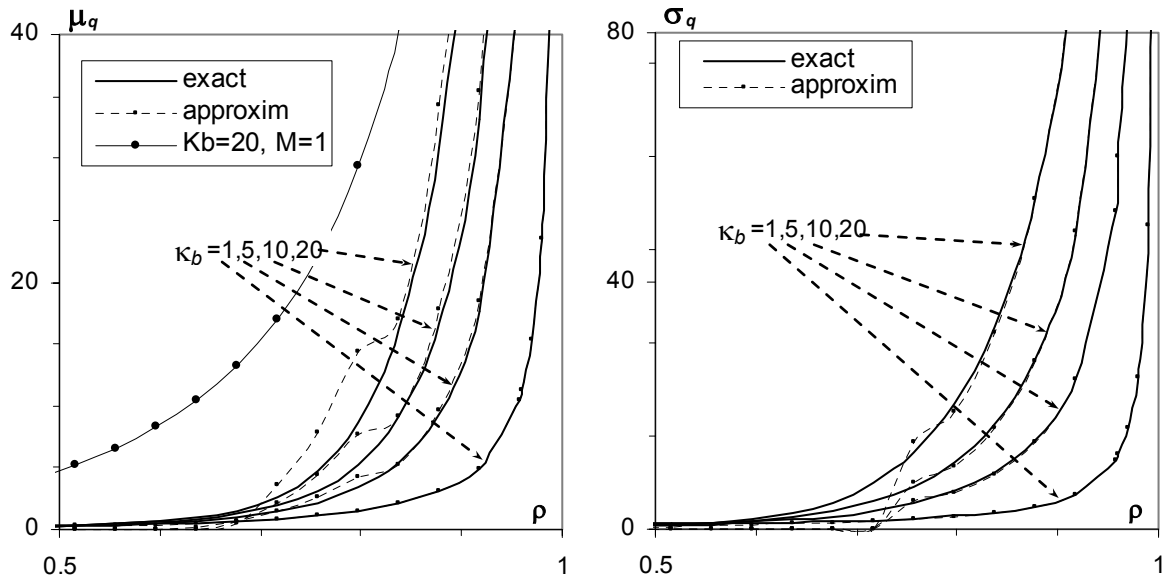


Figure III-8 : moments of the queue content versus κ_s; c=1, N=16, κ_b=5

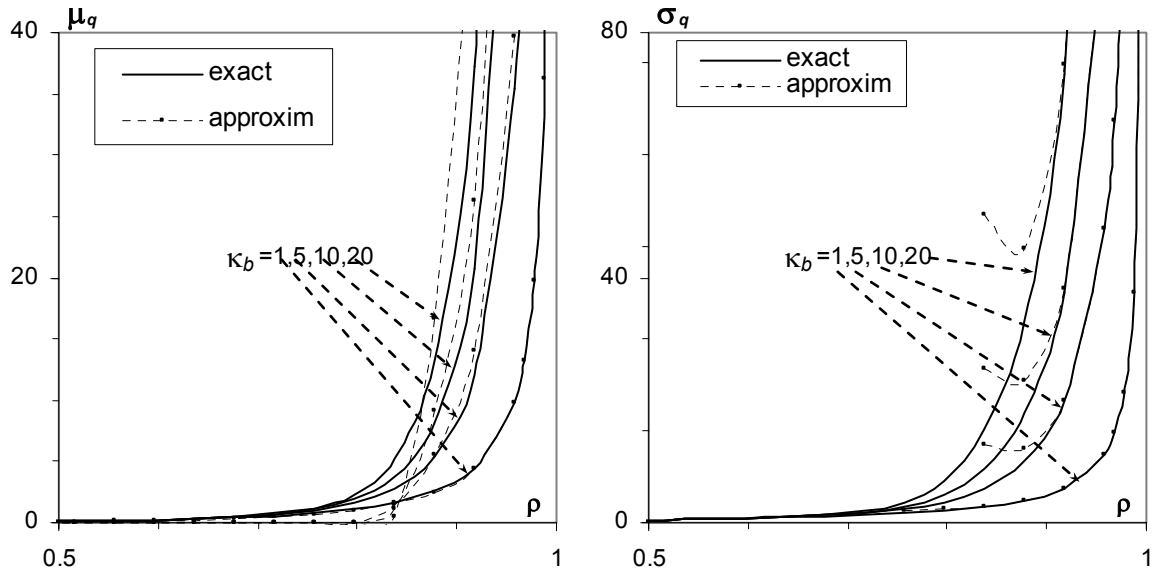
computed in the i.i.d. case by the burst factor κ_b. This once again demonstrates that – in addition to the load ρ – the burst factor κ_b is the defining quantity in the calculations concerning the moments of the queue (and system) content.

The latter conclusion is exemplified by the results shown in Figs. III-8a,b, where we have plotted the mean (a) and stdv (b) of the queue content versus the skew factor κ_s, for κ_b=5, N=16, c=1, and various values of the load ρ as indicated. We see that there exists a negligible (for low ρ) to mild (for high ρ) dependence of these results on the specific value of κ_s, and



(a) mean (b) stdv

Figure III-9 : exact and approximate moments of the queue content, versus ρ; c=1, N=16, $\kappa_s=\kappa_b$



(a) mean (b) stdv

Figure III-10 : exact and approximate moments of the queue content, versus ρ; c=4, N=64, $\kappa_s=\kappa_b$

although not shown here, further examination of the numerical data has led to the ascertainment that this dependence weakens as κ_b increases (for instance for κ_b=20, this dependence on the skew factor κ_s becomes as good as nonexistent).

In the examples that were considered up to now, the exact values of the boundary probabilities were calculated. In the remainder of this section, we make an effort to assess the accuracy of the results if we invoke an approximation such as (III.21a,b) for the boundary probabilities. In Figs III-9a,b one can compare the exact and approximate value of μ_q (a) and

σ_q (b), that have both been plotted versus ρ for N=16, c=1, and values of κ_s=κ_b as indicated. We observe that the approximations that we obtain, both for the mean and stdv of the queue content, are actually quite good, at least when the values of μ_q and σ_q are not too low. In addition and equally important, the accuracy of the results is not influenced by the value of κ_b, again under the condition that ρ is not too low. Note that the irregularities, or 'bumps', that we observe in the curves for the approximations, are due the search algorithm for M, based on criterion (III.22), that switches to a higher value of M at a certain point. In Fig. III-9a, in case of κ_b=20, we have also added the approximation that one obtains for μ_q when applying a fixed value M=1 in the approximation algorithm for the boundary probabilities. These results clearly show that the accuracy is indeed considerably enhanced when values of M larger than 1 are adopted (for ρ sufficiently close to 1, a typical value for M that results from (III.22) would be M=4). Nonetheless – although this can not be clearly viewed on these figures – analysis of the numerical data has confirmed that regardless of the value of M (and κ_b), the proposed approximation converges to the exact results for $\rho \to 1$, since the heavy-load behaviour discussed in Section III.3.5.3 then becomes predominant, and the terms in the expressions for μ_q and σ_q that contain the boundary probabilities become negligible.

Not unexpectedly, in view of the comments that were already made at the end of Section III.3.3, we must conclude that the accuracy of the proposed approximation somewhat deteriorates in the multi-server case, the results of which are shown in Figs. III-10a,b for c=4, N=64, and identical values as above for the traffic parameters that characterise the sources. Still, the above remark that the approximation converges to the exact values for sufficiently high values of ρ remains valid under these circumstances.

III.3.6 tail behaviour of the buffer content distribution

It has been stated on a number of occasions (see e.g. [118], [125], [184], [243], [244], [264], [265], [266], [271], [272], [274], among others) that, also for a buffer with a D-BMAP (or some continuous-time variant that resembles it), an accurate approximation of the tail distribution of the buffer content can be extracted from a dominant-pole approximation, similar to the approach that was followed in the i.i.d. case as discussed in Section II.3.3. Whereas this is true if the threshold T_h – the parameter in $\Pr[X > T_h]$ for some drv X – tends to infinity, the circumstances in a 'real' environment are not always such that T_h is large enough to merit such an approach. We will show by means of some numerical examples that, depending on the D-BMAP traffic parameters, the validity of this assumption is sometimes debatable, and needs to be refined in order to be generally applicable. This, in an almost natural way, leads to the multiple-pole approximation technique presented next, which corresponds to considering a weighted sum of multiple geometric terms in the approximation for the buffer content (see also [14]).

III.3.6.1 the multiple-poles tail approximation

We demonstrate in Appendix A that an approximation for the tail distribution of a drv can be obtained through a series expansion of the corresponding pgf around its poles. If we consider the queue or system content in a buffer with a D-BMAP, then, generally speaking, it no longer suffices to merely adopt a single dominant pole approximation, and multiple poles of the pgf need to be taken into account. The premise that we make is that, in virtually all cases that are of interest for us, the poles that are predominant in the tail behaviour of these drvs can be found on the positive real axis in the region $]1,\mathcal{R}[$, with $\mathcal{R}>1$ the radius of convergence of the pgm $\mathbf{Q}(z)$ as defined in Appendix D (note that the moduli of these poles must perforce be larger than 1, since such a pgf is then a bounded function within the complex unit disk $\{z\in\mathbb{C}:|z|\leq 1\}$).

In Appendix D we show that potential singularities of the pgf $Q(z)$ (that describes the steady-state queue content) that are inherently connected to the decomposition of the pgm $\mathbf{Q}(z)$, such as the branch points of its eigenvalues, are in fact removable singularities of $Q(z)$. Consequently, the non-removable poles of $Q(z)$ (or $S(z)$) will be the zeroes of the denominators in the right hand side of expression (III.25) (or (III.24)) for this pgf, and hence are solutions of the equation $z^c = E_{\bar{\boldsymbol{m}}}(z)$. Then, for any value of $\bar{\boldsymbol{m}}$, each one of these equations may have either 0, 1, or even more than 1, solution(s) in the region $]1,\mathcal{R}[$. Let us define the subset $\Omega_{o,\bar{\boldsymbol{a}}}$ of $\Omega_{\bar{\boldsymbol{a}}}$ (i.e., $\Omega_{o,\bar{\boldsymbol{a}}}\subseteq\Omega_{\bar{\boldsymbol{a}}}$), as the set of values of $\bar{\boldsymbol{m}}$ for which at least 1 such a solution exists. We will then denote by $z_{\bar{\boldsymbol{m}}}$ the smallest of these solutions, for each value of $\bar{\boldsymbol{m}}$. Hence, the dominant poles of the pgf $Q(z)$ (or $S(z)$) are the (real) solutions larger than 1 with the smallest modulus, of each of the equations

$$z_{o,\bar{\boldsymbol{m}}}^c = E_{\bar{\boldsymbol{m}}}\left(z_{o,\bar{\boldsymbol{m}}}\right) \equiv \prod_{j=1}^{L} \lambda_j \left(z_{o,\bar{\boldsymbol{m}}}\right)^{m_j} \;;\; z_{o,\bar{\boldsymbol{m}}} \in]1,\mathcal{R}[\;,\; \bar{\boldsymbol{m}} \in \Omega_{o,\bar{\boldsymbol{a}}} \quad . \tag{III.32a}$$

In the subsequent section we show that the smallest of these poles will be found for $\bar{\boldsymbol{m}} = \bar{\boldsymbol{N}}$, i.e., is the solution of $z^c=\lambda_1(z)^N$ on the positive real axis with modulus larger than 1, and the existence of a solution for each of the above equations will be commented upon as well.

In a quite similar way as in the i.i.d. case (e.g. expressions (II.31b,c) and (II.32)), the series expansion around these poles of the pgf $Q(z)$ then yields the following approximate expressions for the tail ccdf of the queue and system content :

$$\Pr[q>T_h] = \Pr[s>T_h+c] \cong \sum_{\bar{\boldsymbol{m}}\in\Omega_{o,\bar{\boldsymbol{a}}}} \frac{\zeta_{\bar{\boldsymbol{m}}} z_{o,\bar{\boldsymbol{m}}}^{-T_h}}{z_{o,\bar{\boldsymbol{m}}}-1} \quad , \text{ for large enough } T_h \quad , \tag{III.32b}$$

with

$$\zeta_{\bar{\boldsymbol{m}}} \triangleq \lim_{z \to z_{\mathrm{o},\bar{\boldsymbol{m}}}} \left(1 - \frac{z}{z_{\mathrm{o},\bar{\boldsymbol{m}}}}\right) Q(z) = -\frac{c(1-\rho)\left(z_{\mathrm{o},\bar{\boldsymbol{m}}} - 1\right)}{c z_{\mathrm{o},\bar{\boldsymbol{m}}}^{c} - z_{\mathrm{o},\bar{\boldsymbol{m}}} E'_{\mathrm{o},\bar{\boldsymbol{m}}}\left(z_{\mathrm{o},\bar{\boldsymbol{m}}}\right)} \Psi_{\bar{\boldsymbol{m}}}\left(z_{\mathrm{o},\bar{\boldsymbol{m}}}\right) \quad . \tag{III.32c}$$

Note that one does not necessarily needs to calculate the solution $z_{\mathrm{o},\bar{\boldsymbol{m}}}$ of (III.32a) for each and every value of $\bar{\boldsymbol{m}} \in \Omega_{\mathrm{o},\bar{\boldsymbol{a}}}$. More often than not, it is sufficient to start from the dominant-pole contribution, i.e., $\bar{\boldsymbol{m}} = \bar{\boldsymbol{N}}$, and then, by gradually decreasing the first component m_1 of $\bar{\boldsymbol{m}}$ and increase the remaining component(s) accordingly, calculate subsequent terms in the sum over $\bar{\boldsymbol{m}}$ in the right-hand side of (III.32b) until some stopping criterion is met. This criterion could e.g. express that the sum of the terms that correspond to the current value of m_1, is negligible compared to the value of the total intermediate sum that has been obtained thus far.

In addition, we once more highlight the fact that the explicit calculation of the boundary probabilities can be avoided by applying an approximation such as (III.21a,b) for the quantity $\Psi_{\bar{\boldsymbol{m}}}(z)$ in the above expression, which produces closed-form formulae for the constants $\zeta_{\bar{\boldsymbol{m}}}$ that are very efficient from a computational point-of-view. These expressions are particularly appealing in the single-server case c=1, while still being quite accurate, as will be demonstrated in Section III.3.6.3.

III.3.6.2 in search of multiple real dominant poles

First, since $\mathbf{Q}(z)$ is a matrix with positive entries for all $z \in]1,\mathcal{R}[$, it has one real and positive eigenvalue that *exceeds* the moduli of all other eigenvalues (e.g. [164]) for these values of z; this eigenvalue is the PF-eigenvalue $\lambda_1(z)$. Also, remember that $\lambda_1(1)=1$, $\lambda'_1(1)=c\rho/N<1$. Hence, if we consider increasing values of z in the interval $z \in]1,\mathcal{R}[$, and take into account the inequalities

$$E_{\bar{\boldsymbol{N}}}(z) = \lambda_1(z)^N > \left|E_{\bar{\boldsymbol{m}}}(z)\right| \ , \ \forall \bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}} \setminus \bar{\boldsymbol{N}} \ ; \ z \in]1,\mathcal{R}[\quad ,$$

then this explains why, if any of the equations (III.32a) has a solution for $z \in]1,\mathcal{R}[$, the smallest one will be found for $\bar{\boldsymbol{m}} = \bar{\boldsymbol{N}}$. Furthermore, using similar arguments as in [118], where a continuous-time MAP is considered, one can prove that the PF-eigenvalue of $\mathbf{Q}(z)$ is a *strictly increasing* and *convex* function for $z \in]1,\mathcal{R}[$. Hence, equation (III.32a) with $\bar{\boldsymbol{m}} = \bar{\boldsymbol{N}}$ will have a *unique* solution in this region if (see also condition (C.4) in Appendix C)

$$\lim_{\substack{z \to \mathcal{R} \\ <}} \lambda_1(z)^N \Big/ z^c > 1 \quad ,$$

a requirement that we assume to be satisfied from now on. Observe that the smallest pole $z_{\bar{N}}$ of $Q(z)$ and $S(z)$ that we thus obtain also designates the region of convergence of these pgfs.

Making any other claim on the existence of real solutions of equation (III.32a) for a general pgm $\mathbf{Q}(z)$ proves to be close to being impossible. Observe for instance that, unless the number of states in the modulating Markov chain equals 2, we can, in general, not even be assured of the fact that the non-PF-eigenvalues of $\mathbf{Q}(z)$ are real-valued functions for $z \in]1,\mathcal{R}[$. This, however, usually does turn out to be the case for the traffic scenarios that are of interest for us, and can be clarified as follows. If we apply our models in a telecommunications environment, it is often so that the elementary time unit that we have adopted (i.e., 1 slot) is of the same order of magnitude of a packet's transmission time, and is much smaller compared to the average residence time in any given state (which typically corresponds to the mean length of a session, talk spurt, video stream, silent period, …). Consequently, this implies that the values of the diagonal elements of $\mathbf{Q}(1)$ lie relatively close to 1, i.e., for each $1 \le i \le L$,

$$p_{ii} \cong 1 \; ; \; p_{ij} \cong 0 \; , \; 1 \le j \le L \quad \text{and} \quad j \neq i \quad .$$

One can then readily deduce from the characteristic equation that, in particular for values of z on the positive real axis, the eigenvalues of $\mathbf{Q}(z)$ converge to the diagonal elements of this matrix, i.e., $\lambda(z) \rightarrow p_{ii}G_{ii}(z)$, and this correspondence becomes more pronounced as p_{ii} is closer to 1. This does by no means prove, but gives a strong hint that the eigenvalues of $\mathbf{Q}(z)$ are indeed real-valued functions for values of $z \in]1,\mathcal{R}[$ under these particular circumstances.

Let us illustrate this by a small example. We consider a 3×3 MMBP arrival process, where the bgfs $G_{ij}(z)$ are of the form (III.6), with transition probabilities and arrival rates given by the 3×3 matrix

$$\mathbf{Q}(1) = \begin{bmatrix} 1-1/50 & 1/100 & 1/100 \\ 1/20 & 1-1/10 & 1/20 \\ 1/240 & 1/240 & 1\text{-}1/120 \end{bmatrix} \quad , \begin{cases} \gamma_{j1} \equiv \gamma_1 = 0.1 \\ \gamma_{j2} \equiv \gamma_2 = 0.4 \\ \gamma_{j3} \equiv \gamma_3 = 0 \end{cases} \quad ; \quad 1 \le j \le 3 \quad .$$

A source that is characterised by these parameters generates an average of 0.05 packets per slot; also, the radius of convergence of this $\mathbf{Q}(z)$ equals $+\infty$ in such a situation. In Fig. III-11, we have plotted the values of $\lambda_j(z)$, $1 \le j \le 3$, as functions of $z > 1$. Two observations become apparent from this figure : first, the eigenvalues are indeed real-valued functions for real values of $z > 1$, for the reasons explained above. Also, as z increases, the eigenvalues rapidly converge to a similar behaviour as the diagonal elements $p_{ii}G_{ii}(z)$ of $\mathbf{Q}(z)$, which are also shown on this figure as functions of z. Therefore, since $\lambda_1(z)$ and $\lambda_2(z)$ evolve to monotonically and linearly increasing functions of z, while $\lambda_3(z)$ becomes constant for increasing values of $z > 1$, it should be obvious that, for each value of

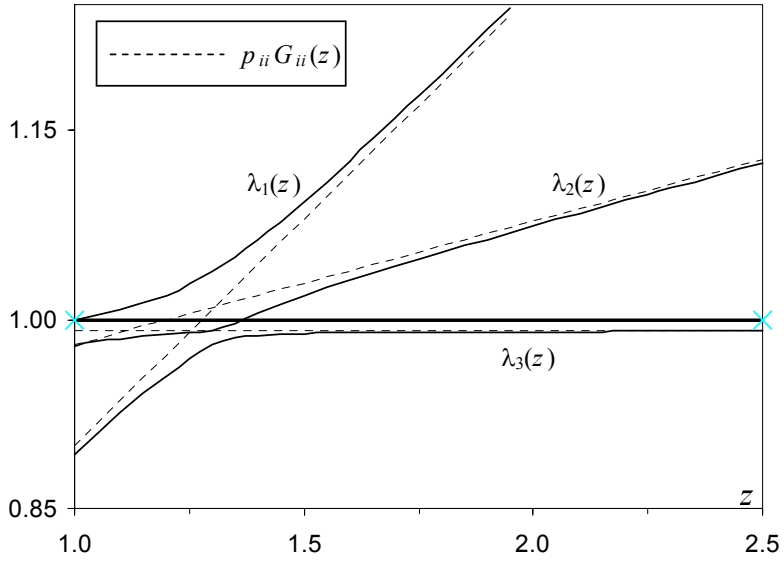


Figure III-11 : eigenvalues of **Q**(*z*), versus *z*

$\bar{\boldsymbol{m}} = [(N - m_2 - m_3) \quad m_2 \quad m_3]^T$, equation (III.32a) has exactly one solution for $z \in]1,\infty[$ if $N - m_3 > c$, since for these values $\bar{\boldsymbol{m}}$, the condition

$$\lim_{\substack{z \to \mathcal{R} \\ <}} E_{\bar{\boldsymbol{m}}}(z) \big/ z^c > 1 \quad ,$$

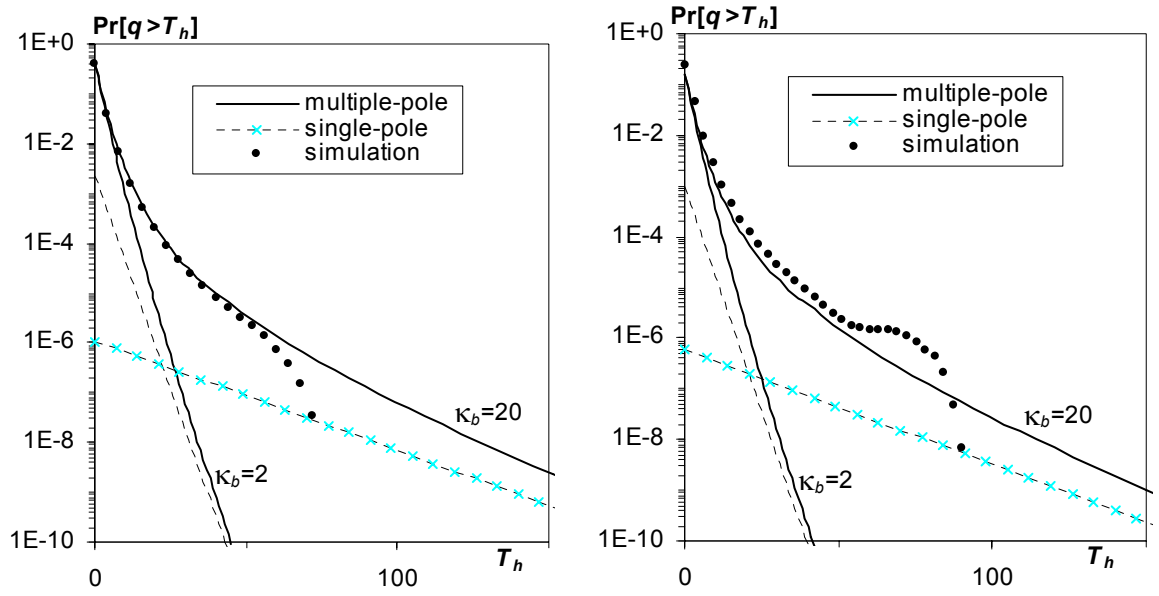
is fulfilled. Therefore, the requirement $N - m_3 > c$ completely defines the set $\Omega_{0,\bar{\boldsymbol{a}}}$ in (III.32b,c).

The above example gives a brief outline on how to deal with the calculation of the dominant poles of the pgfs $Q(z)$ or $S(z)$, provided that the eigenvalues of **Q**(z) are real valued functions on the positive real axis. If such is not the case for all eigenvalues, it is a minor step to adapt the foregoing formulae in order to include the complex solution with the smallest modulus of each of the equations (III.32a), and taking into account that, if $z_{0,\bar{\boldsymbol{m}}}$ is such a solution for some value of $\bar{\boldsymbol{m}}$, then its complex conjugate $(z_{0,\bar{\boldsymbol{m}}})^*$ will be a solution as well.

III.3.6.3 numerical examples

In view of the multiple-pole approximation approach presented in the previous sections, there are two potential sources of inaccuracies that we may need to deal with when we calculate the tail distribution of the queue (or system) content according to the procedure that is outlined therein. To highlight the efficacy of this method, we primarily focus on the 2-state IBP packet arrival process discussed in Section III.3.5.4.

First, numerical errors may be introduced by adopting the multiple-pole series expansion (III.32a-c) with the exact values of the boundary probabilities. The accuracy of this approach



(a) c=1 (b) c=4

Figure III-12 : exact and approximate results for Pr[q>T_h], versus T_h; ρ=0.7; N=80, κ_s=κ_b

is illustrated in Figs. III-12a,b, where we compare the single-pole and multiple-pole approximation technique for the ccdf of the queue content Pr[q>T_h], versus T_h; we consider the single-server (a) and multi-server (b) case with N=80, ρ=0.7, and a burst factor of κ_b=2 and 20 respectively (with κ_s=κ_b from now on, unless mentioned otherwise). For a bursty arrival pattern with κ_b=20, we have also added the exact values of Pr[q>T_h], obtained through simulation. These curves clearly reveal that there is a very good correspondence between the multiple-pole tail approximation and the exact simulated values (at least in the probability region where the latter data points are reliable). In the multi-server case there is a small discrepancy between the two curves, which can be explained by the fact that we have neglected the non-real solutions of (III.32a) for c>1, situated in the positive half-plane – with modulus larger than 1 – that may yield a minor contribution to (III.32b) if taken into account. However, the small difference between the simulated and approximate results indicates that only a marginal gain is to be expected from a numerical procedure that includes these additional terms; moreover, the correspondence is excellent in the single-server case c=1. Also, we need to conclude that a single-pole approximation by no means produces accurate predictions for the ccdf of the queue content, in particular when the traffic becomes more bursty, and these observations justify the multiple-pole approximation technique that was introduced.

A second source of numerical inaccuracies is rooted in the introduction of approximation (III.21a,b) for the boundary probabilities. The parameter M-1 that plays a role in this approach – and denotes the length of a series of slots with zero arrivals, that precede a tagged slot with $s \leq j$ and its predecessor with $e \leq j$ (e.g. Fig. III-2) – will be determined from condition

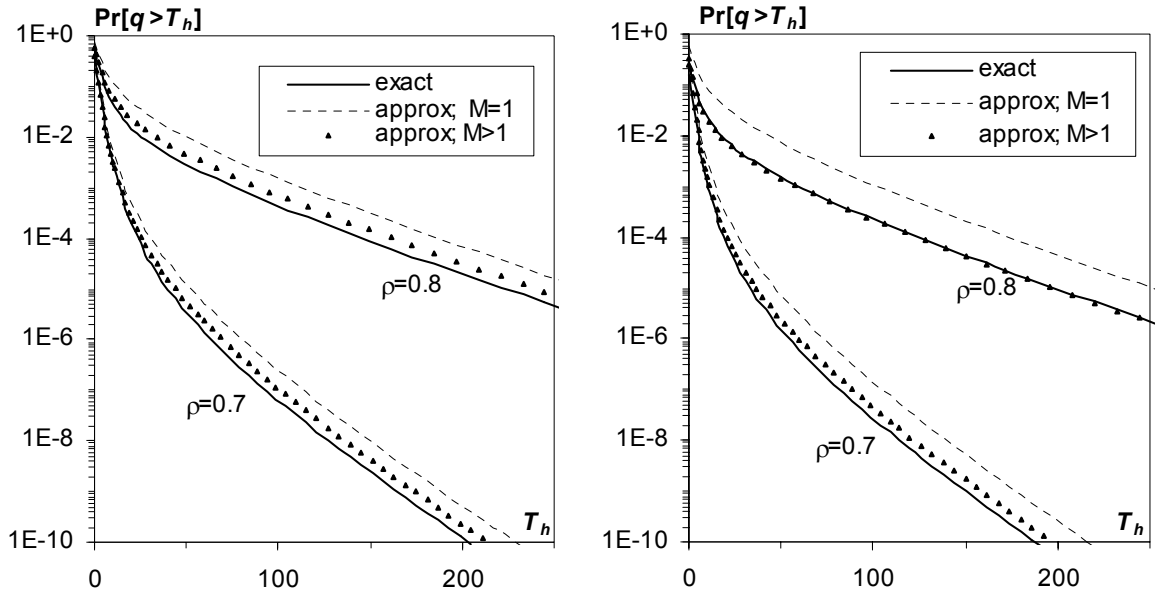


Figure III-13 : exact and approximate results for Pr[q>T_h], versus T_h; N=80, κ_s=κ_b=20

(III.22) if not set equal to M=1. Figs. III-13a,b show the results for the multiple-pole approximation for the queue content ccdf if the exact values of the boundary probabilities are plugged into our formulae, and compare it with the approximate results that we obtain for M=1 and M>1 (i.e., deduced from (III.22)). The traffic parameters are set as N=80, ρ=0.7, κ_b=20, and we once more consider the single-server (a) and multi-server (b) case respectively. Obviously, whereas the approximation with M=1 already gives a more than decent prediction of the queue content ccdf that is conservative (i.e., worst-case), the accuracy of the approximation is clearly enhanced for M>1. Even more so than when computing the moments of the queue content, these results highlight the efficacy of the approach that was followed for approximating the boundary probabilities, and will be further illustrated in the next couples of figures.

In Figs. III-14a,b we have plotted Pr[q>50] as a function of the burst factor κ_b for c=1 (a) and 4 (b), with N=80 and values of the load ρ as indicated. Again, the comparison of the multiple pole approximation when using the exact values of the boundary probabilities, and the one that is deduced from their approximation (with M>1), leads to the ascertainment that the approximate technique yields a highly accurate and efficient way of predicting the 'exact' results, both in the single-server and multi-server scenario. In addition, the queue content ccdf Pr[q>T_h], for fixed T_h, is a strictly increasing function of the burst factor κ_b as expected, but 'flattens out' and converges to a constant value as κ_b becomes larger and larger (note that the moments of the queue and system content become infinitely large for $\kappa_b \to \infty$). Determining this limit may prove to be an interesting topic for future research, since increasing values of the burst factor corresponds to the variance of the total number of packet

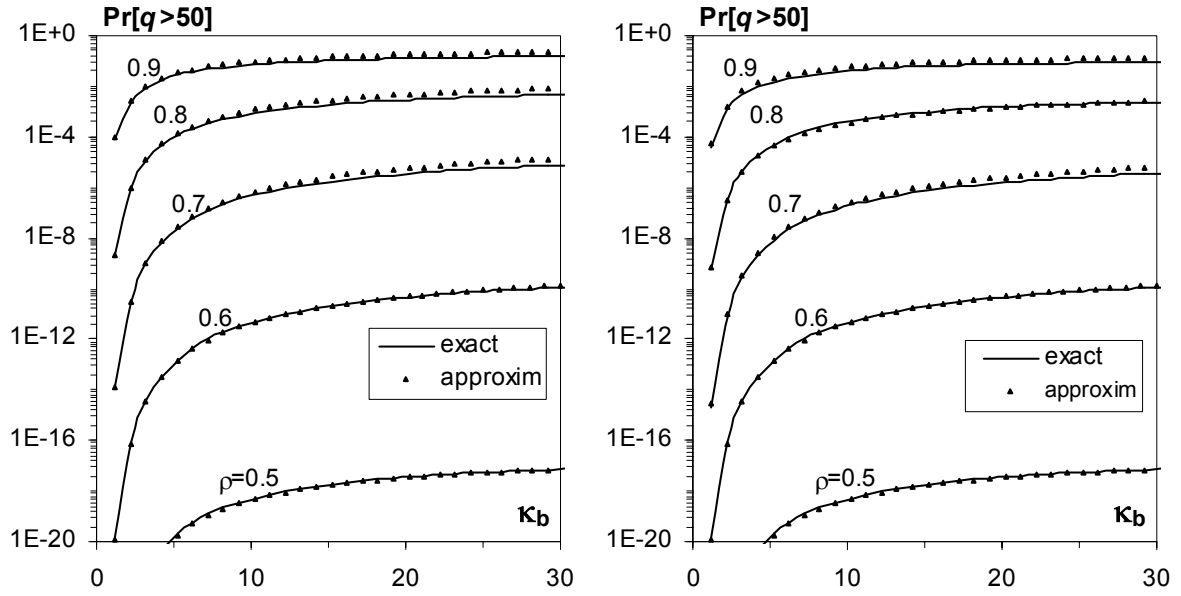
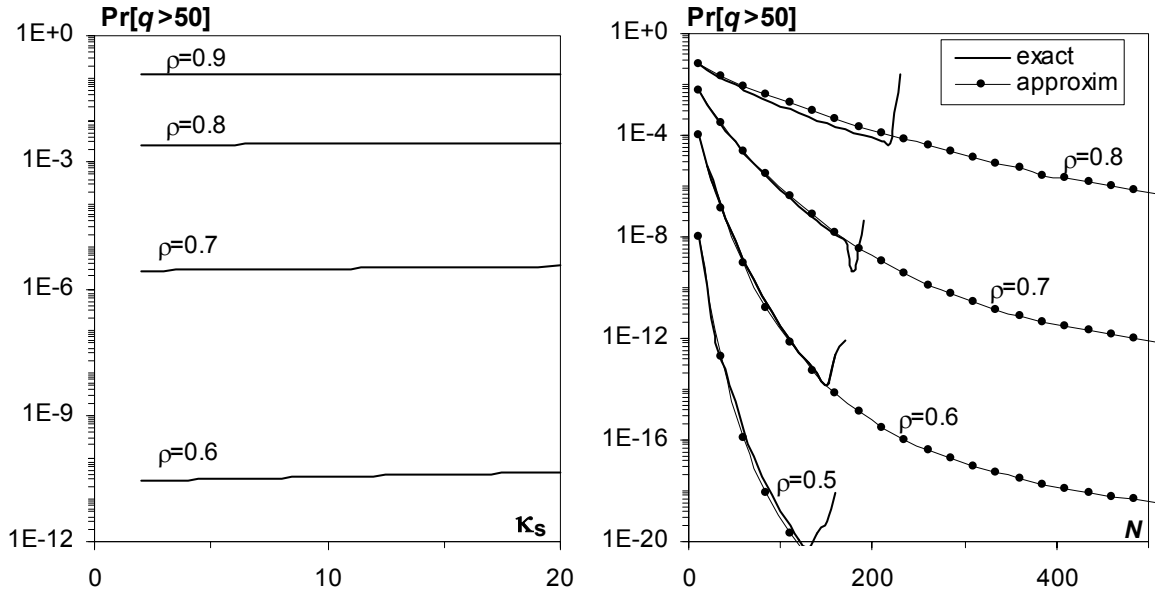


Figure III-14 : Pr[q>50] versus κ_b; N=80, κ_s=κ_b

arrivals during a large time interval that tends to infinity, which is, in fact, related to topics such as long-range dependent packet arrival processes ([168], [225]). Results for performance measures such as the steady-state distribution of the buffer content and/or packet sojourn time are notoriously difficult to obtain under such traffic conditions, and these observations once more demonstrate the efficacy of the computational approach that was followed in the foregoing sections. Apparently, from these and previous figures we may also draw the conclusion that the value of c, the number of servers, does not have a significant impact on the tail behaviour of the queue content, which is the reason why we confine ourselves to the single-server case in the next couple of examples.

Whereas the burst factor has a significant impact on the queue content ccdf, this is not the case as far as the skew factor is concerned, as can be deduced from Fig. III-15a, where we have plotted Pr[q>50] versus κ_s for fixed values of ρ, c, N, and the burst factor as indicated. This emphasises the conclusion that was reached before when considering the moments of the buffer content, namely that the load ρ and the burst factor κ_b are by far the determining parameters in the assessment of the buffer performance.

The necessity of being able to identify sufficiently accurate approximations for the boundary probabilities that are easy to calculate is illustrated once more in Fig. II-15b, where we have plotted Pr[q>50] versus the number of sources N, for c=1, κ_s=κ_b=20, and values of the load ρ as indicated. We again observe a very narrow correspondence between the results that adopt the approximate values for the boundary probabilities, and the curves that are produced by using the exact values of these quantities, at least for those parameter values where the latter are calculable. Indeed, we observe that for increasing values of N, the 'exact'



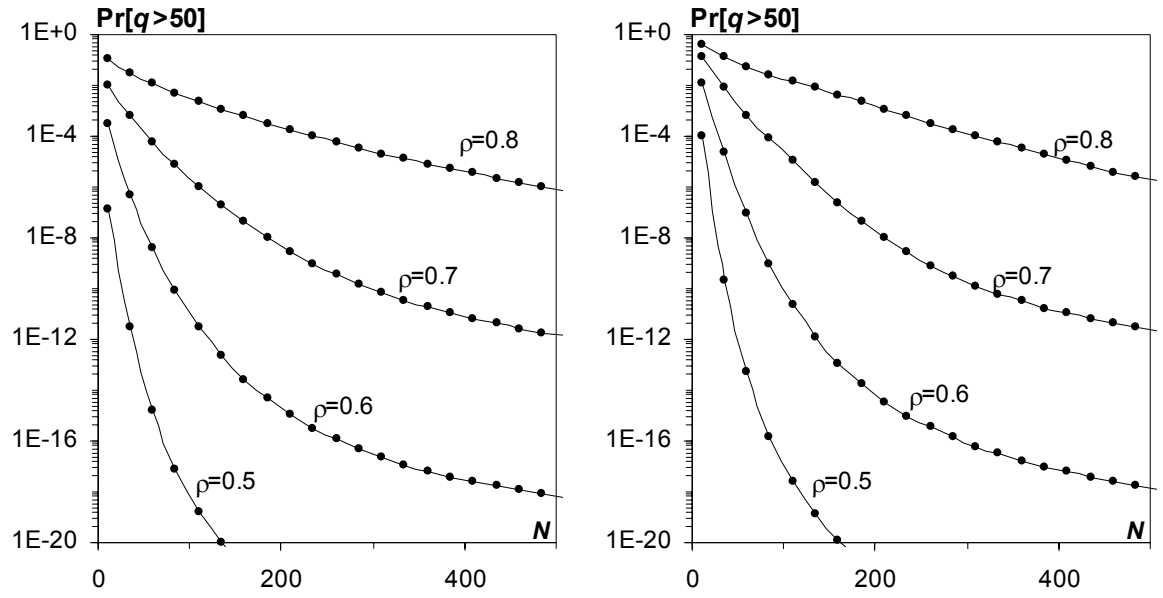
(a) versus κ_s; N=80 (b) versus N; $\kappa_s=\kappa_b$

Figure III-15 : Pr[q>50] for various values of ρ; c=1, κ_b=20

solution method breaks down, due to numerical problems that arise in the computations. These are due to a combination of two phenomena : on the one hand, the increasing size of the set of linear equations for the boundary probabilities that needs to be solved, and on the other hand, a huge – and increasing – difference of several orders of magnitude of the values that these boundary probabilities can assume. Fortunately, the proposed 'approximate' methodology does not suffer from these drawbacks, thereby illustrating its usefulness and efficacy. From this figure we also need to conclude that increasing the number of sources for constant κ_b and κ_s (and ρ) alleviates the burstiness of the aggregate arrival process (significantly, depending on the value of ρ) since the system performance becomes better, which may be a bit counterintuitive, because the opposite effect occurs when the arrival process is i.i.d. From expressions (III.31a,b) one can deduce that doubling the value of N (roughly) halves the values of the arrival rate γ, but (roughly) doubles the average residence time in both the states as well. Apparently, the lower arrival rate in state S_1 preponderates compared to the longer residence times.

Finally, in Figs. III-16a,b we consider similar system parameters as in Fig. III-15b, and a traffic scenario where the arrival process in state S_1 is no longer a Bernoulli process, but a 'clumped' arrival process described by the pgf (II.30a), with a clumping factor equal to κ_c=1 (a) and 5 (b) respectively. For such an arrival pattern, the transition probabilities p_{12} and p_{21} are related to κ_c and κ_b as

$$p_{21}^{-1} = \frac{(\kappa_b - 1)(\kappa_c + \pi_2\gamma)}{2\pi_1\pi_2\gamma} + \frac{1}{\pi_1} \; ; \; p_{12}^{-1} = \frac{\pi_1}{\pi_2} p_{21}^{-1} \quad ,$$



(a) κ_c=1 (b) κ_c=5

Figure III-16 : Pr[q>50] versus N; c=1, κ_b=20, π_1=0.5

where we will set $\pi_1=\pi_2=0.5$; since the skew factor merely has a marginal impact on the system performance, so will the specific value of π_1 for a fixed value of κ_b (this has been checked and confirmed by a wide range of numerical examples, not shown here). Since p_{12}^{-1} and p_{21}^{-1} represent the average residence times in states S_1 and S_2 respectively, these expressions indicate that these quantities are proportional to the values of both κ_b and κ_c. Comparison of these three figures reveals that, as the number of packet arrivals per slot becomes more 'bursty' (i.e., from a Bernoulli, over a Poisson, towards a clumped packet arrival process), there is indeed a significant performance degradation, especially when the number of traffic sources is low. As the number of sources increases however, this performance loss is alleviated, and becomes close to being indiscernible for high values of N. This once again highlights the moderating effect of an increase of the number of traffic sources on the system performance, for fixed ρ and κ_b.

III.4 The packet sojourn times

In this section, we analyse the time that a packet spends in the multi-server output buffer with a correlated arrival process, for a FCFS queuing discipline. By means of an appropriate use of the system equations and the (joint) probability generating functions that can be derived thereof, we will show that a surprisingly simple relation exists between the (steady-state distributions of) the packet sojourn time on the one hand, and the buffer occupancy on the other hand, independent of the specific type of correlations that is imbedded in the packet arrival process. The derivations presented here are principally based on [9] and [54].

III.4.1 the steady-state pgf of the packet waiting time and delay

Adopting the notational conventions of Section II.4, we can reiterate the derivations that led to relation (II.41), to show that the steady-state pgf $W(z)$ that describes the waiting time of a randomly chosen packet satisfies the relation

$$W\left(z^c\right)=\frac{1}{c}\sum_{n=0}^{c-1}\frac{1-z^{-c}}{1-\left(\eta^n z\right)^{-1}}R\left(z\eta^n\right)=\frac{1}{c}\sum_{n=0}^{c-1}\frac{1-z^{-c}}{1-\left(\eta^n z\right)^{-1}}\mathbf{E}\left[\left(z\eta^n\right)^{q^*+f}\right] \quad , \tag{III.33}$$

with $r \triangleq q^*+f$, where the drv q^* denotes the queue content at the beginning of an arbitrary slot during which a randomly chosen packet enters the output buffer, and f the number of packets that have entered the buffer during the same slot and are positioned in the queue before the tagged one.

Under the assumption of an i.i.d. arrival process, it was argued in Section II.4.1 that *i*) the drvs q^* and f are statistically independent, and *ii*) the drv q^* is identically distributed as the drv q, which represents the queue content at the beginning of an arbitrary slot. Evidently, these statements are no longer valid if the arrival process is a correlated one – such as the D-BMAP under consideration – and we therefore need to adapt and refine our approach. If we denote by e^* the total number of packet arrivals during the tagged packet's arrival slot, then invoking the law of total probability, we obviously can write

$$\Pr[q^*=i,f=j]=\sum_{k=j+1}^{\infty}\Pr[f=j|q^*=i,e^*=k]\Pr[q^*=i,e^*=k] \quad .$$

First, since the tagged packet has been chosen in a random fashion among a total of e^* packets that have arrived during its arrival slot, its position among those packets is subject to a uniform distribution, yielding

$$\Pr[f=j|q^*=i,e^*=k]=\frac{1}{k} \;;\; 0\le j\le k-1 \quad .$$

Moreover, by means of a analogous reasoning as in Section II.4.1, one can show that the joint pmf of the couple (q^*,e^*) can be expressed in terms of the joint pmf of (q,e) (where q represents the queue content at the beginning of an arbitrary slot, and e the total number of arrivals during *the same* slot) as

$$\Pr[q^*=i,e^*=k]=\frac{k\Pr[q=i,e=k]}{c\rho} \quad ,$$

where we let $c\rho$ represent the average number of packet arrival during an arbitrary slot, as before. Hence, if we define $H(z,x)$ as the steady-state joint pgf of the pair of drvs (q^*,f), and

$G(z,x)$ as the joint pgf of the pair (q,e) respectively, then the combination of the three previous expressions leads to

$$H(z,x) \triangleq \mathbf{E}\left[z^{q^*}x^f\right] = \frac{1}{c\rho}\sum_{i=0}^{\infty} z^i \sum_{j=0}^{\infty} x^j \sum_{k=j+1}^{\infty} \Pr[q=i,e=k]$$

$$= \frac{1}{c\rho}\sum_{i=0}^{\infty} z^i \sum_{k=0}^{\infty} \frac{x^k-1}{x-1}\Pr[q=i,e=k] = \frac{G(z,x)-G(z,1)}{c\rho(x-1)} \quad .$$

On the other hand, the evolution of the queue content at consecutive slot boundaries is governed by the system equation

$$q_{n+1} = (q_n + e_n - c)^+ \quad ,$$

which immediately follows from (I.1a,b). If we z-transform this expression assuming steady-state conditions, and take into account that $Q(z) \equiv G(z,1)$, we find that

$$z^c G(z,1) = G(z,z) + (z-1)\sum_{j=0}^{c-1} z^j s_c(j) \quad ,$$

where $s_c(j)$ represents the cdf of the system content s, as before, and where we have made use of the property

$$\lim_{n\to\infty} \Pr[q_n + e_n \le j] = \lim_{n\to\infty} \Pr[s_{n+1} \le j] = s_c(j) \quad .$$

Consequently, since $R(z)$, the steady-state pgf of q^*+f, equals $H(z,z)$, we obtain

$$R(z) = \frac{z^c-1}{c\rho(z-1)}Q(z) - \frac{1}{c\rho}\sum_{j=0}^{c-1} z^j s_c(j) \quad ,$$

and the quantities that appear in the right-hand side of this expression merely pertain to the queue and system content at the beginning of an arbitrary slot, and do not require any knowledge whatsoever that is related to the packet arrival process, apart from its mean rate.

Using this result, the previous expression for the steady-state pgf $W(z)$ of the packet waiting time can then be transformed into

$$W\left(z^c\right) = \frac{1}{\rho}\sum_{n=0}^{c-1} U_c\left(\left(z\eta^n\right)^{-1}\right)\left(U_c\left(z\eta^n\right)Q\left(z\eta^n\right) - \frac{1}{c}\sum_{j=0}^{c-1}\left(z\eta^n\right)^{j+1} s_c(j)\right) \quad ,$$

where $U_c(z)$ represents the pgf of a drv that is uniformly distributed between 1 and c, as in (II.42a). The second term in the right-hand side can now be rewritten as

$$\frac{1}{c}\sum_{n=0}^{c-1}\sum_{j=0}^{c-1}U_c\left(\left(z\eta^n\right)^{-1}\right)\left(z\eta^n\right)^{j+1}s_c(j)=\frac{1}{c^2}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}\sum_{n=0}^{c-1}\left(z\eta^n\right)^{j-i}s_c(j) \quad .$$

In view of the identity (II.40) and the summation boundaries for i and j, the sum over n is nonzero only if $i=j$ (for each value of j). We thus find

$$\frac{1}{c}\sum_{n=0}^{c-1}\sum_{j=0}^{c-1}U_c\left(\left(z\eta^n\right)^{-1}\right)\left(z\eta^n\right)^{j+1}s_c(j)=\frac{1}{c}\sum_{j=0}^{c-1}s_c(j)=1-\rho \quad ,$$

where we have invoked the normalisation condition (II.8), which remains valid in case of a correlated packet arrival process. We thus finally obtain

$$\boxed{W\left(z^c\right)=-\frac{1-\rho}{\rho}+\frac{1}{\rho}\sum_{n=0}^{c-1}U_c\left(\left(z\eta^n\right)^{-1}\right)U_c\left(z\eta^n\right)Q\left(z\eta^n\right)} \quad , \qquad \text{(III.34a)}$$

an identity that relates (the steady-state pgfs of) the waiting time of an arbitrary packet on the one hand, and the queue content at the start of an arbitrary slot on the other hand. This relation can be viewed as an extension, and indeed refinement, of (II.42a) that was valid for an i.i.d. packet arrival process, to the case of a non-independent packet arrival stream.

A similar relation can be established for the steady state pgf $S(z)$ that describes the system content at the beginning of an arbitrary slot, based on the relation

$$Q(z)=z^{-c}S(z)+z^{-c}(z-1)\sum_{j=0}^{c-1}z^j s_c(j) \quad .$$

Taking into account that $(x\text{-}1)U_c(x)=x(x^c-1)/c$, then observe that we can also deduce that

$$\begin{aligned}\sum_{n=0}^{c-1}\sum_{j=0}^{c-1}(x_n-1)U_c\left(x_n^{-1}\right)U_c(x_n)x_n^j s_c(j)\Big|_{x_n=z\eta^n} &= \frac{1}{c^2}\left(z^c-1\right)\sum_{j=0}^{c-1}s_c(j)\sum_{n=0}^{c-1}\sum_{i=0}^{c-1}\left(z\eta^n\right)^{j-i}\\ &=\left(z^c-1\right)(1-\rho) \quad .\end{aligned}$$

Therefore, because the steady-state pgf $D(z)$ of the packet delay satisfies $D(z)=zW(z)$, then combining the above expressions results in

$$\boxed{D\left(z^c\right)=-\frac{1-\rho}{\rho}+\frac{1}{\rho}\sum_{n=0}^{c-1}U_c\left(\left(z\eta^n\right)^{-1}\right)U_c\left(z\eta^n\right)S\left(z\eta^n\right)} \quad . \qquad \text{(III.34b)}$$

An conspicuous feature of the expressions (III.34a,b) is that the steady-state pgfs of the packet waiting time and the queue content on the one hand, and the packet delay and the system content on the other hand, satisfy the same formal relationship. Another crucial and

remarkable observation is that we have not in any way relied on the specific nature of the correlation between the packet arrivals during successive slots throughout the derivations in this section. Therefore, since the relations (III.34a,b) contain no reference whatsoever with respect to the specifics of the arrival process (apart from the mean $c\rho$), it will be valid *no matter what* the packet arrival pattern looks like (i.e., i.i.d., Markovian, long-range dependent, …). Consequently, once performance measures – such as mean, variance and (tail) distribution – of the queue (or system) content have been determined, it is relatively straightforward to establish expressions for the mean, variance and (tail) distribution of the packet waiting time (or delay), as will be demonstrated next.

III.4.2 (tail) distribution of the packet sojourn time

From expression (III.34a) that relates the steady-state pgfs of the queue content and packet waiting time, we can also extract a corresponding relationship between their respective pmfs. Adopting similar techniques as above, this expression can be transformed into

$$c\rho W\left(z^c\right) = -c(1-\rho) + \frac{1}{c}\sum_{n=0}^{c-1}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}\Pr[q=k]\left(z\eta^n\right)^{k+j-i} \quad .$$

In view of (II.40), the sum over n will be equal to c if $k+j-i$ is a multiple of c, and 0 otherwise. Hence, since $k+j-i \geq -(c-1)$ for all values of i,j,k, the above expression can be rewritten as

$$c\rho W\left(z^c\right) = -c(1-\rho) + \sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}\delta(mc+i-j-k)\Pr[q=k]z^{mc} \quad .$$

For $m>0$, there is exactly one value of k for which $k=mc+i-j$; for $m=0$, this is only the case if $i \geq j$. We thus find

$$c\rho W\left(z^c\right) = \sum_{m=1}^{\infty}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}\Pr[q=mc+i-j]z^{mc} + \sum_{i=0}^{c-1}\sum_{j=0}^{i}\Pr[q=i-j] - c(1-\rho) \quad .$$

Consequently, if we identify the coefficients of z^m in both hand sides of this equation, we finally obtain

$$c\rho \cdot \Pr[w=m] = \begin{cases} \sum_{k=0}^{c-1}(c-k)\Pr[q=k] - c(1-\rho) \ , & m=0 \\ \sum_{k=0}^{c-1}\left(k\cdot\Pr[q=(m-1)c+k] + (c-k)\cdot\Pr[q=mc+k]\right) \ , & m>0 \end{cases} \quad . \qquad \text{(III.35a)}$$

From a similar treatment of expression (III.34b) for $D(z^c)$, combined with the normalisation condition (II.8) (which leads to the value $\Pr[d=0]=0$ as expected, since the packet delay is at least one slot), we deduce that we may immediately write

$$c\rho \cdot \Pr[d=m] = \sum_{k=0}^{c-1} (k \cdot \Pr[s=(m-1)c+k] + (c-k) \cdot \Pr[s=mc+k]) \ , \ m>0 \quad . \qquad \text{(III.35b)}$$

Note that, due to $\Pr[s=i+c]=\Pr[q=i]$ for $i>0$, we also find that $\Pr[w=m]=\Pr[d=m+1]$, for $m>0$, as expected. These results lead to the following expressions for the ccdfs of the packet delay and waiting time, in terms of the ccdfs (and pmfs) of the queue and system content

$$\begin{aligned} c\rho \cdot \Pr[w>T_h] &= c\Pr[q>c(T_h+1)] + \sum_{k=1}^{c} k \cdot \Pr[q=cT_h+k] \\ c\rho \cdot \Pr[d>T_h] &= c\Pr[s>c(T_h+1)] + \sum_{k=1}^{c} k \cdot \Pr[s=cT_h+k] \end{aligned} \quad ; \ T_h \geq 0 \quad . \qquad \text{(III.36a)}$$

In order to briefly verify our calculations, observe that it is straightforward to check that these pmfs and ccdfs of the packet waiting time and delay are normalised; if we, for instance, insert the value $T_h=0$ in the above expression for $\Pr[d>T_h]$, we obtain

$$\begin{aligned} c\rho\Pr[d>0] &= c\Pr[s \geq c] + \sum_{k=0}^{c-1} k \cdot \Pr[s=k] = c\Pr[s \geq 0] - \sum_{k=0}^{c-1} (c-k) \cdot \Pr[s=k] \\ &= c - c(1-\rho) = c\rho \ , \end{aligned}$$

which indeed yields that $\Pr[d>0]=1$.

Finally, if we adopt a multiple-pole approximation as in Section III.3.6.1 to calculate the probabilities in the right-hand side of (III.36a), then after some derivations we obtain the following expressions for the asymptotic tail behaviour of the queue and system content :

$$\begin{aligned} \Pr[w>T_h] &\cong \frac{1}{\rho} \sum_{\bar{m} \in \Omega_{o,\bar{a}}} \frac{\zeta_{\bar{m}} U_c(z_{o,\bar{m}})}{z_{o,\bar{m}} - 1} z_{o,\bar{m}}^{-c(T_h+1)} \\ \Pr[d>T_h] &\cong \frac{1}{\rho} \sum_{\bar{m} \in \Omega_{o,\bar{a}}} \frac{\zeta_{\bar{m}} U_c(z_{o,\bar{m}})}{z_{o,\bar{m}} - 1} z_{o,\bar{m}}^{-cT_h} \end{aligned} \quad . \qquad \text{(III.36b)}$$

These expressions indicate that $z_{o,\bar{m}}^{c}$ are the poles of the steady-state pgfs $W(z)$ and $D(z)$ outside the complex unit circle that determine the tail behaviour of the associated drvs.

III.4.3 mean and variance of the packet sojourn time

If we take the first derivative with respect to z for $z=1$ of (III.34a), and take into account that $U_c(\eta^n)=0$ unless $n=0$, we can establish the following relation

$$c\rho W'(1) = -\frac{c+1}{2} + \frac{c+1}{2} + Q'(1) \quad ,$$

which immediately yields, exploiting the formal resemblance of (III.34a,b),

$$\begin{aligned} c\rho \cdot \mu_w &= \mu_q \\ c\rho \cdot \mu_d &= \mu_s \end{aligned} \quad , \tag{III.37}$$

which shows that the pair of random variables (w,q) on the one hand and (d,s) on the other hand indeed satisfies Little's theorem, as expected. Note that these relations can also be easily deduced by multiplication by m of (III.35a,b) and taking the sum for all $m>0$, e.g.,

$$\begin{aligned} c\rho \cdot \mu_w &= \sum_{m=1}^{\infty} m \sum_{k=0}^{c-1} (k \cdot \Pr[q=(m-1)c+k] + (c-k) \cdot \Pr[q=mc+k]) \\ &= \sum_{m=1}^{\infty} \sum_{k=0}^{c-1} ((m+1)k + m(c-k)) \cdot \Pr[q=mc+k] = \sum_{k=0}^{c-1} k \cdot \Pr[q=k] \equiv \mu_q \quad . \end{aligned}$$

The computations concerning the variance of the packet waiting time and delay are a bit more involved. First note that with definition (II.14) of the $\mathcal{V}[\cdot]$ operator, we may write

$$\mathcal{V}\left[\rho W\left(z^c\right)\right] = c^2\rho\left(\sigma_w^2 + \mu_w^2\right) - c^2\rho^2\mu_w^2 = c^2\rho\left(\sigma_w^2 + \mu_w^2\right) - \mu_q^2 \quad .$$

Again relying on the property that $U_c(\eta^n)=0$ unless $n=0$, then, in view of expression (III.34a) and the computational rules (II.15a,b), we obtain

$$\begin{aligned} c^2\rho\left(\sigma_w^2 + \mu_w^2\right) &= \mu_q^2 + \mathcal{V}\left[U_c\left(z^{-1}\right)U_c(z)Q(z)\right] + \mathcal{V}\left[\sum_{n=1}^{c-1} U_c\left(\left(z\eta^n\right)^{-1}\right)U_c\left(z\eta^n\right)Q\left(z\eta^n\right)\right] \\ &= \sigma_q^2 + \mu_q^2 + \frac{c^2-1}{6} - 2\sum_{n=1}^{c-1} Q\left(\eta^n\right)\mathcal{M}\left[U_c\left(\left(z\eta^n\right)^{-1}\right)U_c\left(z\eta^n\right)\right] \quad , \end{aligned}$$

where we have also applied the property that the variance of a drv that is uniformly distributed between 1 and c equals $(c^2-1)/12$. From the equality

$$\frac{1}{c}\sum_{j=0}^{c-1} j\eta^{nj} = \frac{1}{\eta^n - 1} \quad , \; 1 \le n \le c-1 \quad ,$$

and the formal resemblance between (III.34a,b) we thus find that

$$\sigma_w^2+\mu_w^2=\frac{\sigma_q^2+\mu_q^2}{c^2\rho}+\frac{c^2-1}{6c^2\rho}-\frac{2}{c^2\rho}\sum_{n=1}^{c-1}\left|\frac{1}{\eta^n-1}\right|^2 Q\left(\eta^n\right)$$
$$\sigma_d^2+\mu_d^2=\frac{\sigma_s^2+\mu_s^2}{c^2\rho}+\frac{c^2-1}{6c^2\rho}-\frac{2}{c^2\rho}\sum_{n=1}^{c-1}\left|\frac{1}{\eta^n-1}\right|^2 S\left(\eta^n\right) \quad . \qquad \text{(III.38a)}$$

Thus, in order to calculate the variance of the packet waiting time and delay, we need to evaluate the pgfs $Q(z)$ and $S(z)$ in the c-1 complex solutions of $z^c=1$ different from $z=1$, in addition to calculating the first two moments of the queue and system content. Depending on the explicit form of $Q(z)$ and $S(z)$, calculating $Q(\eta^n)$ and $S(\eta^n)$ may or may not be feasible, and a computational task that we might wish to avoid. For that purpose, we derive a tight upper and lower bound for the complex sum(s) in the right-hand side of expression (III.38a).

To that extent, let us define

$$S1 \triangleq \sum_{n=1}^{c-1}\left|\frac{1}{\eta^n-1}\right|^2 Q\left(\eta^n\right)=\frac{1}{c^2}\sum_{m=0}^{\infty}\sum_{k=0}^{c-1}\Pr[q=mc+k]\sum_{n=0}^{c-1}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}ij\eta^{n(k+j-i)}-\frac{(c-1)^2}{4} \quad ,$$

where the last term in the right-hand side reflects the term for $n=0$ in the preceding sum. Let us first calculate

$$S2 \triangleq \sum_{n=0}^{c-1}\sum_{j=0}^{c-1}\sum_{i=0}^{c-1}ij\eta^{n(k+j-i)}=\sum_{n=0}^{c-1}\sum_{j=0}^{c-k-1}\sum_{i=0}^{c-1}ij\eta^{n(k+j-i)}+\sum_{n=0}^{c-1}\sum_{j=c-k}^{c-1}\sum_{i=0}^{c-1}ij\eta^{n(k+j-i)} \quad .$$

Since $0\le k+j\le c$-1 for the first term in the right-hand side, $i=k+j$ is the only value of i for which the sum over n in this term is nonzero, while $i=k+j-c$ is the only value of i for which the sum over n in the second term is nonzero, due to $c\le k+j\le 2(c$-1). Therefore, we can derive that

$$S2=c\sum_{j=0}^{c-k-1}j(k+j)+c\sum_{j=c-k}^{c-1}j(k+j-c)$$
$$=\frac{c^2(c-1)(2c-1)}{6}-k\frac{c^2(c-1)}{2} \quad .$$

Inserting the minimum and maximum value of k in the above expression, i.e., 0 and c-1 respectively, results in a lower and upper bound for $S2$, and hence, $S1$,

$$\frac{c^2-1}{12}-\frac{(c-1)^2}{2}\le S1\le\frac{c^2-1}{12} \quad .$$

In view of the definition of $S1$, inserting these bounds in (III.38a) then eventually yields

$$0 \le \sigma_w^2 + \mu_w^2 - \frac{\sigma_q^2 + \mu_q^2}{c^2\rho} \le \frac{(c-1)^2}{c^2\rho}$$
$$0 \le \sigma_d^2 + \mu_d^2 - \frac{\sigma_s^2 + \mu_s^2}{c^2\rho} \le \frac{(c-1)^2}{c^2\rho} \quad , \qquad \text{(III.38b)}$$

Note that these inequalities produce an exact relationship between the first two moments of the queue (system) content and the packet waiting time (delay) in the single-server case c=1. These bounds indicate that the margins between which the waiting time (delay) variance can vary, are actually very tight for relatively high loads (i.e., in the order of 1 slot^2).

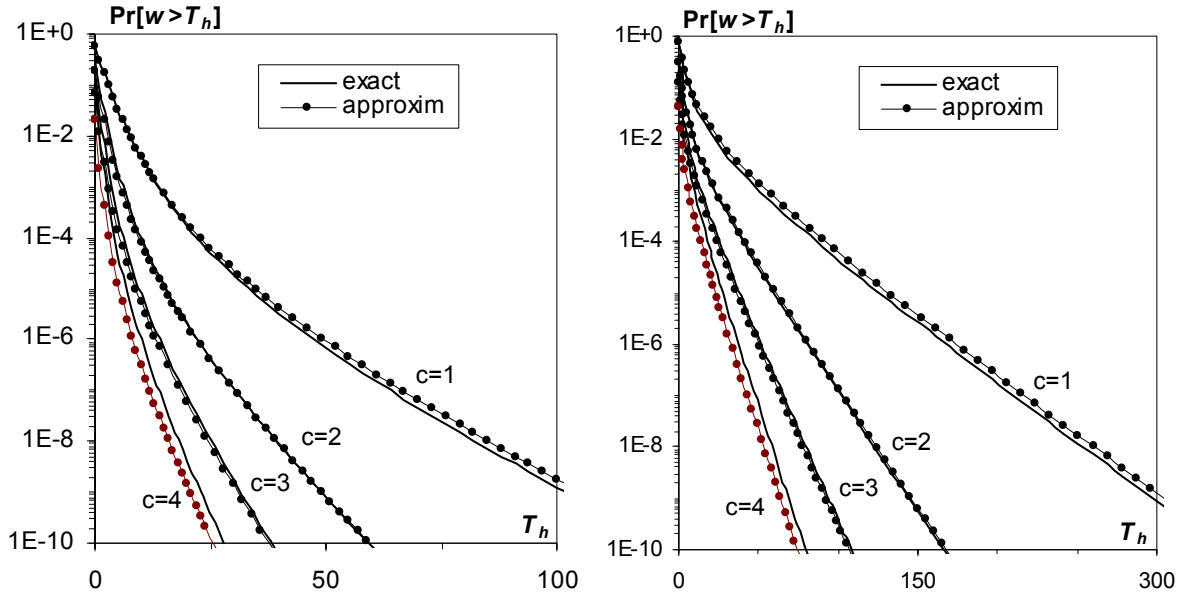
Finally, we point out that an approximate expression for the variance of the waiting time (delay) in terms of the first two moments of the queue (system) content can be derived as well from these result, by substituting the 'average' value of k, being (c-1)/2, in the latter expression for $S2$, which leads to the following result :

$$\sigma_w^2 + \mu_w^2 \cong \frac{\sigma_q^2 + \mu_q^2}{c^2\rho} + \frac{(c-1)^2}{2c^2\rho}$$
$$\sigma_d^2 + \mu_d^2 \cong \frac{\sigma_s^2 + \mu_s^2}{c^2\rho} + \frac{(c-1)^2}{2c^2\rho} \quad . \qquad \text{(III.38c)}$$

III.4.4 numerical examples

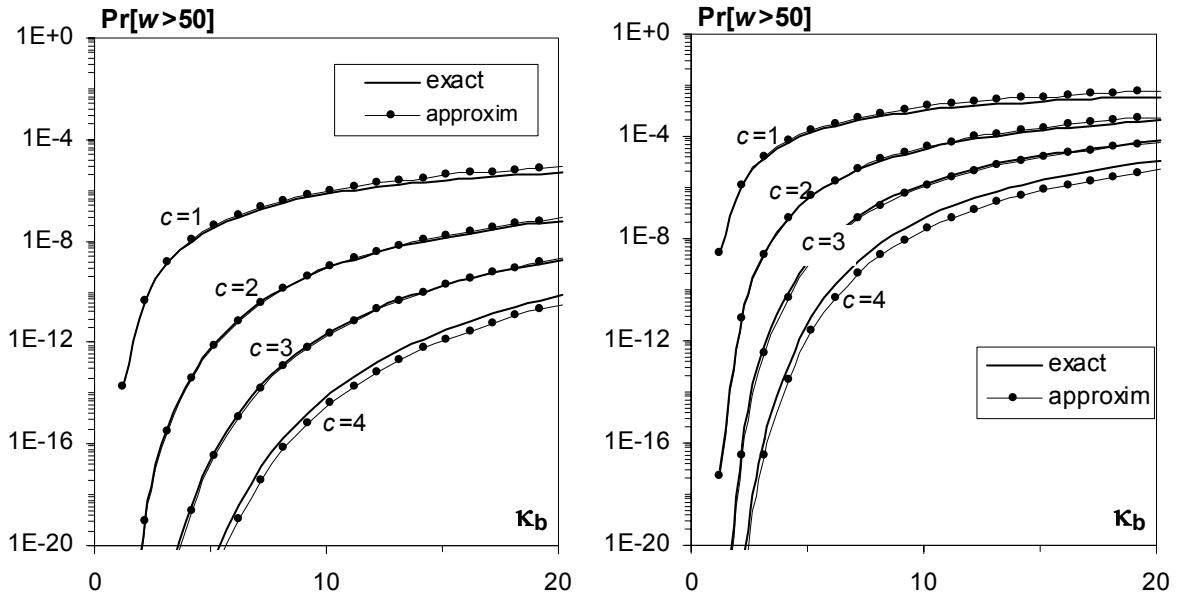
We present just a few numerical examples, in order to illustrate the effect of multi-server queuing systems on the packet waiting time (and delay). When considering the case of i.i.d. arrivals, we concluded that the main benefit of adopting multi-server queues stems from the reduction of the waiting times and delays, rather than the queue sizes. A similar assessment can be made in case of correlated packet arrival processes such as a D-BMAP, as can be viewed on Figs. III-17a,b, where, for c=1,2,3,4, we have depicted the waiting time ccdf Pr[w>T_h] versus T_h, for values of N=80, κ_s=κ_b=10, and ρ=0.7 (a) and 0.8 (b) respectively. From these figures we can deduce that the 10^{-X} quantile of the packet waiting time is (roughly) inversely proportional to the value of c, irrespective of the value the load ρ and/or X. In addition, as before, these curves illustrate the strong correspondence between the 'exact' results that have been obtained from using the exact values of the boundary probabilities in expression (III.36b) for the tail behaviour, and the approximate values that are computed by using approximation (III.21b) for these quantities. Whereas the asymptotic tail behaviour of the waiting time is slightly overestimated by the proposed approximation for c=1, the opposite becomes true as c increases.

The accuracy of this approximation is further illustrated in Fig. III-18a,b as well, where we show Pr[w>50] versus the burst factor κ_b, for N=80, ρ=0.7 (a) and 0.8 (b) respectively, and



(a) ρ=0.7 (b) ρ=0.8

Figure III-17 : Pr[w>T_h] versus T_h; N=80, κ_s=κ_b=10



(a) ρ=0.7 (b) ρ=0.8

Figure III-18 : Pr[w>50] versus κ_b; N=80, κ_s=κ_b

c=1,2,3,4, which basically yields similar conclusions as above with respect to the accuracy of the approximation technique. These results also highlight the strong decline of the values of Pr[w>T_h] as c increases, as expected in view of Figs. III-17a,b. Finally, similar to the behaviour that we observed when we studied the tail distribution of the queue content, we must conclude that Pr[w>T_h] converges to a horizontal asymptote as κ_b increases, an intriguing phenomenon that merits further investigation, but falls beyond our current scope.

III.5 Heterogeneous packet arrival processes

Up to now we have primarily focused on queuing systems whereby the N sources that generate packets have the same (stochastic) characteristics, and hence, can be described by a single D-BMAP. Depending on the communications environment this may not always be a realistic assumption, and it therefore stands to reason to extend the previous results to the case of heterogeneous, meaning non-identical, packet arrival processes that originate from the respective sources (see also [14]).

III.5.1 a heterogeneous D-BMAP

Consider therefore an output buffer that is fed by N independent sources, as before. These sources are grouped into K distinct classes or types, according to their traffic characteristics, each class comprising N_k, $1 \le k \le K$, independent and identical sources. A source that belongs to class k is modelled as an L_k-state D-BMAP of the same type as described in Section III.2.1, where the states that can be visited by each of these sources are now labelled $S_{i,k}$, $1 \le i \le L_k$, $1 \le k \le K$.

Similar to the case of a homogeneous D-BMAP, the so-called probability generating matrices $\mathbf{Q}_k(z)$ will be central to the calculations in our analysis, and these matrices can be written as (e.g. (III.1))

$$\mathbf{Q}_k(z) \triangleq \left[q_{ij,k}(z)\right] = \begin{bmatrix} p_{11,k}G_{11,k}(z) & p_{12,k}G_{12,k}(z) & \cdots & p_{1L_k,k}G_{1L_k,k}(z) \\ p_{21,k}G_{21,k}(z) & p_{22,k}G_{22,k}(z) & \cdots & p_{2L_k,k}G_{2L_k,k}(z) \\ \vdots & \vdots & \ddots & \vdots \\ p_{L_k1,k}G_{L_k1,k}(z) & p_{L_k2,k}G_{L_k2,k}(z) & \cdots & p_{L_kL_k,k}G_{L_kL_k,k}(z) \end{bmatrix},$$

where $p_{ij,k}$ represents the one-step transition probabilities that a source of class k transits from state $S_{i,k}$ to state $S_{j,k}$ from the previous to the current slot, and where the bgfs $G_{ij,k}(z)$ describe the number of packet arrivals that are being generated in the current slot if such a transition occurs for a source of type k; $1 \le i,j \le L_k$ and $1 \le k \le K$.

We can now define $a_{l,n,k}$ as the drv that equals the total number of inlets of class k that visit state S_l during slot n, and denote by the vector $\overline{\boldsymbol{a}}_{k,n} \equiv [a_{1,k,n} \cdots a_{L_k,k,n}]^T$ the set of drvs that captures the state of the class-k sources during slot n. In view of the previous, these drvs satisfy

$$\sum_{l=1}^{L_k} a_{l,k,n} = N_k \quad ; \quad \sum_{k=1}^{K} N_k = N \quad , \tag{III.39}$$

which expresses that during any slot, each of the N_k sources of type k visits one of the states $S_{l,k}$, $1 \le l \le L_k$, for all $1 \le k \le K$. These considerations thus reveal that the set of vectors of drvs

$\{\overline{\boldsymbol{a}}_{k,n} : 1 \le k \le K\}$ can be adopted to characterise the state of the N sources at the start of successive slots. For increasing n, each set $\overline{\boldsymbol{a}}_{k,n}$ forms a Markov chain with sample space

$$\Omega_{\overline{\boldsymbol{a}}_k} = \left\{ \left[l_{1,k} \cdots l_{L_k,k} \right]^T : l_{i,k} \ge 0 \wedge \sum_{i=1}^{L_k} l_{i,k} = N_k ; \right\} ; \; 1 \le k \le K \quad .$$

Hence, the set (of sets of drvs) $\{\overline{\boldsymbol{a}}_{k,n} : 1 \le k \le K\}$ forms a Markov chain as well, with state space

$$\Omega_{\overline{\boldsymbol{a}}_1 \cdots \overline{\boldsymbol{a}}_K} = \left\{ \Omega_{\overline{\boldsymbol{a}}_1}, \cdots, \Omega_{\overline{\boldsymbol{a}}_K} \right\} \quad , \tag{III.40a}$$

and the cardinality of this set is now equal to

$$\left| \Omega_{\overline{\boldsymbol{a}}_1 \cdots \overline{\boldsymbol{a}}_K} \right| = \prod_{k=1}^{K} \binom{N_k + L_k - 1}{N_k} \quad , \tag{III.40b}$$

which is the amount of different states that can be visited by this Markov chain, that determines the heterogeneous packet arrival process.

Moreover, if we define by $e_{k,n}$ the drv that describes the number of packet arrivals generated by the aggregation of N_k sources of class k, then through a similar reasoning as the one that led to (III.8), we can show that

$$\mathbf{E}\left[z^{e_{k,n}} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k,n}} \right] = \mathbf{E}\left[\left(\mathbf{Q}_k(z) \overline{\boldsymbol{x}}_k \right)^{\overline{\boldsymbol{a}}_{k,n-1}} \right] \quad , \tag{III.41a}$$

with $\overline{\boldsymbol{x}}_k \triangleq \left[x_{1,k} \cdots x_{L_k,k} \right]^T$, $1 \le k \le K$, and where we have adopted an identical notational shorthand convention as defined in equation (III.2). Finally, denoting by the drv e_n the total number of packet arrivals during slot n that originates from the aggregation of N sources, i.e.,

$$e_n \triangleq \sum_{k=1}^{K} e_{k,n} \quad ,$$

then by invoking the independence of sources that belong to different classes, the penultimate expression can be readily extended to

$$\mathbf{E}\left[z^{e_n} \prod_{k=1}^{K} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k,n}} \right] = \mathbf{E}\left[\prod_{k=1}^{K} \left(\mathbf{Q}_k(z) \overline{\boldsymbol{x}}_k \right)^{\overline{\boldsymbol{a}}_{k,n-1}} \right] \quad . \tag{III.41b}$$

As in the homogeneous traffic scenario, this expression is a shorthand matrix notation for an equation that governs the evolution at (i.e., just after) successive slot boundaries of both the packet arrival process and the state of the underlying Markov chain that defines it. This equation will form the basis of the analysis of the multi-server queuing system with a heterogeneous D-BMAP that we currently investigate.

Furthermore, if we define $A_k(\bar{x}_k)$ as the steady-state joint pgf that describes the state of the underlying Markov chain of the sources of type k, then with the aid of the results of Section III.2.2 one readily shows that

$$A_k(\bar{x}_k) \triangleq \lim_{n\to\infty} \mathbf{E}\left[\bar{x}_k^{\bar{a}_{k,n}}\right] = \left(\bar{\pi}_k{}^T \bar{x}_k\right)^{N_k}, \tag{III.42}$$

where $\bar{\pi}_k$ is the stationary vector (with components $\pi_{l,k}$, $1 \le l \le L_k$) that describes the state of a source of type k during an arbitrary slot, which is the solution of

$$\bar{\pi}_k{}^T = \bar{\pi}_k{}^T \mathbf{Q}_k(1) \; ; \; \bar{\pi}_k{}^T \bar{I}_k = \bar{I}_k \quad ,$$

where $\bar{I}_k$ is the $L_k \times 1$ vector whose components are all equal to 1. Hence, the steady-state pgf $E(z)$ associated with the drv e, that describes the (total) number of packet arrivals during an arbitrary slot, is given by

$$E(z) \triangleq \mathbf{E}\left[z^e\right] = \prod_{k=1}^{K} E_k(z) \; ; \; E_k(z) \triangleq A_k\left(\mathbf{Q}_k(z)\bar{I}_k\right) = \left(\bar{\pi}_k^T \mathbf{Q}_k(z)\bar{I}_k\right)^{N_k} \quad . \tag{III.43a}$$

Therefore, a straightforward calculation leads to the following expression for the average number of packet arrivals per slot, denoted by $c\rho$ as before :

$$c\rho \triangleq \mu_e = E'(1) = c\sum_{k=1}^{K} \rho_k \; ; \; c\rho_k \triangleq N_k \sum_{l=1}^{L}\sum_{m=1}^{L} \pi_{l,k}\, p_{lm,k}\, G'_{lm,k}(1) \quad , \tag{III.43b}$$

where ρ_k represents the mean number of packet arrivals per slot and per output link, of type k. In addition, the variance of e can by similarly derived, and is given by

$$\sigma_e^2 = \sum_{k=1}^{K} N_k \sum_{l=1}^{L}\sum_{m=1}^{L} \pi_{l,k}\, p_{lm,k}\, G''_{lm,k}(1) + \sum_{k=1}^{K} c\rho_k \left(1 - \frac{c\rho_k}{N_k}\right) \quad . \tag{III.43c}$$

The third central moment $\mu_{3,e}$ of e can be computed analogously; e.g. (III.11d).

The properties of the heterogeneous D-BMAP described above, for instance with respect to correlation and (in)dependence, readily follow from the results that were presented in Section III.2.3 for the homogeneous case, and will not be explicitly recapitulated at this point; obviously, the aggregated heterogeneous D-BMAP will be an uncorrelated (independent) process if each of its homogeneous components is an uncorrelated (independent) process, and vice versa. Let us just conclude by mentioning that the subsequent analysis will also (partly) rely on the decomposition of the pgms $\mathbf{Q}_k(z)$ into their respective eigenvalues and –vectors, and the remarks with respect to the existence of these quantities that were made in Section III.2.1 and Appendix D naturally remain valid in the current scenario. In the results that

follow hereafter, we will represent by $\mathbf{W}_k(z)$ (with rows $\vec{w}_{i,k}(z)$ and elements $w_{ij,k}(z)$), and $\mathbf{U}_k(z)$ (with columns $\vec{u}_{i,k}(z)$ and elements $u_{ji,k}(z)$), the $L_k \times L_k$ matrices that contain the left and right eigenvectors of $\mathbf{Q}_k(z)$ respectively, while we will denote by $\mathbf{\Lambda}_k(z)$ the $L_k \times L_k$ diagonal matrix that contains the corresponding eigenvalues $\lambda_{i,k}(z)$; $1 \le i,j \le L_k$, $1 \le k \le K$. The eigenvectors of $\mathbf{Q}_k(z)$ are defined by analogous equations as (III.7a-c), and the associated remarks with respect to their properties and existence therefore are unabatedly applicable. The PF-eigenvalue of the pgm of type k, denoted as $\lambda_{1,k}(z)$, $1 \le k \le K$, will play an important part in our analysis as well, and the formulae for the performance indices that are deduced thereof. Note, for instance, that the mean number of packet arrivals per slot of type k satisfies $c\rho_k = N_k \lambda'_{1,k}(1)$, while the burst factor κ_b, which serves as a measure for the burstiness of the heterogeneous D-BMAP and is still defined as in (III.15a), can now be calculated from

$$\kappa_b \sigma_e^2 \triangleq \sum_{k=1}^{K} \kappa_{b,k} \sigma_{e_k}^2 = \sum_{k=1}^{K} N_k \left(\lambda''_{1,k}(1) + \lambda'_{1,k}(1) - \lambda'_{1,k}(1)^2 \right) \equiv \sum_{k=1}^{K} \mathcal{V}\left[\lambda_{1,k}(z)^{N_k} \right] \quad , \qquad \text{(III.44a)}$$

where the burst factor $\kappa_{b,k}$ of the individual traffic classes is defined in the same way as described in Section III.2.4. Similarly, in accordance to (III.15c), the skew factor of the aggregate heterogeneous arrival process is now defined as

$$\kappa_s \mu_{3,e} \triangleq \sum_{k=1}^{K} N_k \left\{ \lambda'''_{1,k}(1) + 3\lambda''_{1,k}(1)\left(1 - \lambda'_{1,k}(1)\right) + \lambda'_{1,k}(1)\left(1 - \lambda'_{1,k}(1)\right)\left(1 - 2\lambda'_{1,k}(1)\right) \right\} \quad . \qquad \text{(III.44b)}$$

It will also be implicitly assumed that each of the pgms $\mathbf{Q}_k(z)$ has a radius of convergence $\mathcal{R}_k$ larger than 1, which among others implies that – for a buffer with infinite storage capacity, c output links, and packet transmission times that equal 1 slot – the requirement $\rho < 1$ is a necessary and sufficient condition for the system to reach a stochastic equilibrium. The radius of convergence $\mathcal{R}$ (>1) that applies for the aggregate heterogeneous packet arrival process will then be given by the minimum of the $\mathcal{R}_k$'s, i.e.,

$$\mathcal{R} \triangleq \min_{1 \le k \le K} \{\mathcal{R}_k\} \quad .$$

III.5.2 analysis of the buffer content

If we let s_n denote the system content at the beginning of slot n, being the total number of packets in the output buffer at this time instant, then the system equation

$$s_{n+1} = (s_n - c)^+ + \sum_{k=1}^{K} e_{k,n} \quad , \qquad \text{(III.45)}$$

governs the evolution of the system content at consecutive slot boundaries. For similar reasons as in Section III.3, we can use the vector $(s_{n+1}, \bar{a}_{1,n},\cdots,\bar{a}_{K,n})$, also called the state vector, to describe the state of this queuing system during consecutive slots, i.e., the system content at the start of slot n+1, and sets of drvs $\bar{a}_{k,n}$, that capture the state of the class-k Markov chain during slot n and steers the number of packet arrivals of type k during that slot, $1\leq k\leq K$. If the condition $\rho<1$ is fulfilled, the system will reach a stochastic equilibrium, and we can define the steady-state joint pgf of the state vector as

$$P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K) \triangleq \lim_{n\to\infty} \mathbf{E}\left[z^{s_n} \prod_{k=1}^{K} \bar{\boldsymbol{x}}_k^{\,\bar{a}_{k,n-1}} \right] = \mathbf{E}\left[z^{s} \prod_{k=1}^{K} \bar{\boldsymbol{x}}_k^{\,\bar{a}_{k}} \right] \quad .$$

Adopting the same analysis techniques as in Section III.3.1, a functional equation for this quantity can be constructed, which can be written as

$$P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K) = z^{-c}\left\{ P_s(z,\mathbf{Q}_1(z)\bar{\boldsymbol{x}}_1,\cdots,\mathbf{Q}_K(z)\bar{\boldsymbol{x}}_K) + R(z,\mathbf{Q}_1(z)\bar{\boldsymbol{x}}_1,\cdots,\mathbf{Q}_K(z)\bar{\boldsymbol{x}}_K) \right\} \quad , \text{(III.46a)}$$

with

$$R(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K) \triangleq (z-1)\sum_{j=0}^{c-1} z^j \sum_{\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K} \left(\prod_{k=1}^{K} \bar{\boldsymbol{x}}_k^{\,\bar{l}_k} \right) \Pr\left[s\leq j, \bar{\boldsymbol{a}}_1=\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{a}}_K=\bar{\boldsymbol{l}}_{\boldsymbol{K}} \right] \quad . \qquad \text{(III.46b)}$$

Observe that the drvs that appear in the right-hand side of the latter expression refer to the system content at the start of an arbitrary slot and the state of the packet arrival process during the preceding slot. Unless indicated otherwise, the sum over $\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K$ (and $\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K$ in the remainder) in these and forthcoming expressions runs over all possible values that belong to the sample space $\Omega_{\bar{\boldsymbol{a}}_1\cdots\bar{\boldsymbol{a}}_K}$.

Deriving a closed-form solution for this functional equation, not unexpectedly, proceeds along analogous lines as in Section III.3.2. First, successive substitutions of the arguments in the last expression for $P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$ allow us to write

$$P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K) = z^{-Hc}\left\{ P_s\left(z,\mathbf{Q}_1(z)^H\bar{\boldsymbol{x}}_1,\cdots,\mathbf{Q}_K(z)^H\bar{\boldsymbol{x}}_K\right) + \sum_{h=1}^{H} R\left(z,\mathbf{Q}_1(z)^h\bar{\boldsymbol{x}}_1,\cdots,\mathbf{Q}_K(z)^h\bar{\boldsymbol{x}}_K\right) \right\} .$$

Next, we must define the functions $F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(z)$ through

$$\prod_{i=1}^{L_k}\left(\sum_{j=1}^{L_k} u_{ij,k}(z)x_{j,k} \right)^{l_{i,k}} \triangleq \sum_{\bar{\boldsymbol{m}}_k\in\Omega_{\bar{\boldsymbol{a}}_k}} F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(z)\bar{\boldsymbol{x}}_k^{\,\bar{\boldsymbol{m}}_k} \quad \Leftrightarrow \quad (\mathbf{U}_k(z)\bar{\boldsymbol{x}}_k)^{\bar{\boldsymbol{l}}_k} \triangleq \sum_{\bar{\boldsymbol{m}}_k} F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(z)\bar{\boldsymbol{x}}_k^{\,\bar{\boldsymbol{m}}_k} \quad .$$

This equation then allows us to write

$$\prod_{k=1}^{K}\left(\mathbf{Q}_k(z)^h\,\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{l}}_k}=\prod_{k=1}^{K}\left(\sum_{\bar{\boldsymbol{m}}_k\in\Omega_{\bar{\boldsymbol{a}}_k}}F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(z)\left(\boldsymbol{\Lambda}_k(z)^h\,\mathbf{W}_k(z)\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}\right)\quad . \tag{III.47}$$

If we let $H\to\infty$ in the latter expression for the joint pgf of the state vector and consider values of z that belong to $\{z\in\mathbb{C}:|z|=1\wedge z\neq 1\}$, then the first term in the right hand side will vanish, while the sum for h in the second term will converge and can be worked out in a similar way as in the homogeneous-arrivals case, eventually leading to

$$\boxed{P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)=c(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)\prod_{k=1}^{K}\left(\boldsymbol{\Lambda}_k(z)\mathbf{W}_k(z)\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}}{z^c-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\quad , \tag{III.48a}$$

and where,

$$\begin{aligned}
&\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)\triangleq\sum_{\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K}\left(\prod_{k=1}^{K}F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(z)\right)\sum_{j=0}^{c-1}z^j\,p\left(j,\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)\\
&p\left(j,\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)\triangleq\frac{1}{c(1-\rho)}\Pr\left[s\le j,\bar{\boldsymbol{a}}_1=\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{a}}_K=\bar{\boldsymbol{l}}_K\right]\\
&E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)\triangleq\prod_{k=1}^{K}\left(\boldsymbol{\Lambda}_k(z)\bar{\boldsymbol{l}}_k\right)^{\bar{\boldsymbol{m}}_k}\equiv\prod_{k=1}^{K}\prod_{i=1}^{L_k}\lambda_{i,k}(z)^{m_{i,k}}\quad .
\end{aligned}\tag{III.48b}$$

Moreover, a similar derivation as the one that led to (III.18b) also yields an expression for the steady-state joint pgf of the queue content at the beginning of a slot, and the state of the D-BMAP during the previous slot

$$\boxed{P_q(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)=c(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)\prod_{k=1}^{K}\left(\mathbf{W}_k(z)\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}}{z^c-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\tag{III.48c}$$

These expressions for the joint pgfs $P_q(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$ and $P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$ that describe the queue and system content at the start of an arbitrary slot and the state of the packet arrival process during the foregoing slot, although a bit involved, still allow us to derive closed-form expressions for performance measures such as the mean and variance of the system and/or queue content that are easy to handle, once the boundary probabilities $p\left(j,\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)$ have been computed. Unfortunately, the number of boundary probabilities now, in general, equals $c\cdot\left|\Omega_{\bar{\boldsymbol{a}}_1\cdots\bar{\boldsymbol{a}}_K}\right|$ which will be too high to be of any practical use in a considerable amount of heterogeneous traffic scenarios. It therefore becomes increasingly important to provide approximations for these unknowns that are both accurate and computable without great additional effort.

III.5.2.1 calculation and approximation of the boundary probabilities

Extending the techniques described in Section III.3.3, the boundary probabilities can, in principle, be calculated by expressing that the zeroes of the denominators of $P_S(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$ and/or $P_q(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$ – which will be denoted by $z_{i,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}$ and are the solutions of

$$z = E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)^{1/c}\,\eta^i \;\; ; \;\; \{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K\}\in\Omega_{\bar{\boldsymbol{a}}_1\cdots\bar{\boldsymbol{a}}_K} \;\; , \; 0\le i\le c-1 \quad , \tag{III.49a}$$

inside the complex unit circle $\{z\in\mathbb{C}:|z|<1\}$ – must also be zeroes of the corresponding numerators. Combined with the normalisation condition, this yields the following set of $c\cdot|\Omega_{\bar{\boldsymbol{a}}_1\cdots\bar{\boldsymbol{a}}_K}|$ linear equations for the same number of boundary probabilities :

$$\begin{aligned} &\sum_{\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K}\sum_{j=0}^{c-1} p\left(j,\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)=1 \\ &\left(z_{i,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}-1\right)\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z_{i,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\right)=0 \;;\; \{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K\}\in\Omega_{\bar{\boldsymbol{a}}_1\cdots\bar{\boldsymbol{a}}_K} \;\; , \; 0\le i\le c-1 \; . \end{aligned} \tag{III.49b}$$

Observe that the latter equation becomes trivial and can be omitted for $z_{0,\bar{\boldsymbol{N}}_1\cdots\bar{\boldsymbol{N}}_K}=1$, where by convention $\bar{\boldsymbol{N}}_k\in\Omega_{\bar{\boldsymbol{a}}_k}$ represents the set of integers $\{m_{1,k}\cdots m_{L_k k}\}$ with $m_{1,k}=N_k$, and $m_{j,k}=0$; $2\le j\le L_k$, $\forall\;\; 1\le k\le K$.

As explained in the homogeneous D-BMAP case, the number of boundary probabilities that we need to calculate, will be reduced if one or more of the states $S_{i,k}$ that can be visited by a source of type k are so-called greedy states; generally speaking however, the size of this set of linear equations quickly becomes very high. As a small example, if we set K=2, N_1=N_2=50, c=4, then this would imply that a set of 10404 linear equations for the same number of unknowns needs to be solved. To make things even worse, the value of the boundary probabilities in such a setting is likely to vary over an extremely wide range of values, which strongly indicates that numerically solving such a set of equations is close to being unfeasible. We therefore resort to finding approximations for these quantities that are sufficiently accurate and easy to handle. Fortunately, it is not difficult to show that the approach of Section III.3.3 can be readily extended to the heterogeneous case, by replacing the event $s\le j$ at the start of a slot by $e\le j$ during the preceding slot, which, in turn, is preceded by a time period of M-1 contiguous slots during which no packets arrivals occur. Also making use of the normalisation condition, this approach results in the following approximation for the functions $\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)$ that appear in the expression for $P(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$:

$$\Psi_{a,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)=C_a\sum_{j=0}^{c-1}z^j\sum_{i=0}^{j}\theta_{i,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)$$

$$\theta_{i,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)\triangleq\frac{1}{i!}\frac{\partial^i}{\partial y^i}\left[\prod_{k=1}^{K}\binom{N_k}{\bar{\boldsymbol{m}}_k}\prod_{l=1}^{L_k}\left(\bar{\boldsymbol{\pi}}_k^T\mathbf{Q}_k(0)^{M-1}\mathbf{Q}_k(y)\bar{\boldsymbol{u}}_{l,k}(z)\right)^{m_{l,k}}\right]_{y=0} \qquad \text{(III.50)}$$

$$C_a^{-1}=\sum_{i=0}^{c-1}(c-i)\frac{1}{i!}\frac{\partial^i}{\partial y^i}\left[\prod_{k=1}^{K}\left(\bar{\boldsymbol{\pi}}_k^T\mathbf{Q}_k(0)^{M-1}\mathbf{Q}_k(y)\bar{\boldsymbol{I}}_k\right)^{N_k}\right]_{y=0},$$

an approximation that is especially easy to handle in the single-server case $c=1$, since the derivatives with respect to y in the latter formulae vanish. In the multi-server case, the tractability of this approximation technique depends on the value of c and the specifics of the pgms $\mathbf{Q}_k(z)$, but remains feasible if this matrices for instance all represent a MMBP arrival process. As in the homogeneous arrivals case, the parameter M can be tuned to be the smallest integer for which

$$C_a^{-1}\le c(1-\rho) \quad .$$

III.5.2.2 pgf, moments and tail distribution of the queue and system content

If we insert $\bar{\boldsymbol{x}}_k=\bar{\boldsymbol{I}}_k$, for all $1\le k\le K$, into the expression (III.48a) for $P_s(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$ and take into account that $\mathbf{W}_k(z)\bar{\boldsymbol{I}}_k=\bar{\boldsymbol{I}}_k$, then we immediately find a closed-form formula for $S(z)$, the steady-state pgf that describes the system content at the beginning of an arbitrary slot,

$$\boxed{S(z)=c(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}{z^c-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)} \quad . \qquad \text{(III.51a)}$$

Analogously, result (III.48c) for $P_q(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$, with $\bar{\boldsymbol{x}}_k=\bar{\boldsymbol{I}}_k$, produces the following expression for the steady-state pgf $Q(z)$ of the queue content at the start of an arbitrary slot :

$$\boxed{Q(z)=\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{c(1-\rho)(z-1)}{z^c-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)} \quad . \qquad \text{(III.51b)}$$

Hence, also in the case of a heterogeneous D-BMAP, there is a striking formal similarity between these two formulae, and expressions (II.13) and (II.10) for $S(z)$ and $Q(z)$ in case of an i.i.d. packet arrival process.

This resemblance will again be fully exploited in our derivations concerning the mean, variance, and tail distribution of the queue and system content. Observe that relations such as

(III.26a,b) remain valid for the $F^{(k)}_{\bar{l}_k \bar{m}_k}(z)$ that were defined in Section III.5.2, and we may thus write, similar to (III.26c)

$$\Psi_{\bar{m}_1 \cdots \bar{m}_K}(1) = \begin{cases} 1 \; ; \; \{\bar{m}_1 \cdots \bar{m}_K\} = \{\bar{N}_1 \cdots \bar{N}_K\} \\ 0 \; ; \; \forall \{\bar{m}_1 \cdots \bar{m}_K\} \in \tilde{\Omega} \end{cases} \; ; \; \tilde{\Omega} \triangleq \Omega_{\bar{a}_1 \cdots \bar{a}_K} \setminus \{\bar{N}_1 \cdots \bar{N}_K\} \quad .$$

Consequently, if we take the first derivative with respect to z of (III.51b) for z=1, then the sum over $\bar{m}_1 \cdots \bar{m}_K$ in the right-hand side yields a nonzero term only if $\{\bar{m}_1 \cdots \bar{m}_K\} = \{\bar{N}_1 \cdots \bar{N}_K\}$, and we obtain for the mean queue content

$$\begin{aligned} \mu_q &= \frac{\kappa_b \sigma_e^2}{2(c-\mu_e)} - \frac{\mu_e}{2} + \mathcal{M}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] - \frac{c-1}{2} \\ \mathcal{M}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] &= \sum_{j=0}^{c-1} \sum_{\bar{l}_1 \cdots \bar{l}_K} \left(j + \sum_{k=1}^{K} \sum_{i=1}^{L} l_{i,k} u'_{i1,k}(1) \right) p\left(j, \bar{l}_1, \cdots, \bar{l}_K\right) \end{aligned} \quad , \qquad \text{(III.52a)}$$

where the burst factor κ_b is computed as in expression (III.44a). The mean system content can be similarly computed, leading to

$$\mu_s = \mu_q + \mu_e \quad . \qquad \text{(III.52b)}$$

Furthermore, an identical approach towards the calculation of the variance of the queue content as in Section III.3.5.1 yields

$$\begin{aligned} \sigma_q^2 &= \frac{\kappa_s \mu_{3,e}}{3(c-\mu_e)} + \left(\frac{\kappa_b \sigma_e^2}{2(c-\mu_e)} \right)^2 - \frac{\kappa_b \sigma_e^2}{2} + \frac{1-(c-\mu_e)^2}{12} \\ &\quad + \mathcal{V}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] + 2 \sum_{\{\bar{m}_1 \cdots \bar{m}_K\} \in \tilde{\Omega}} \frac{c(1-\rho)}{1 - E_{\bar{m}_1 \cdots \bar{m}_K}(1)} \mathcal{M}\left[\Psi_{\bar{m}_1 \cdots \bar{m}_K}\right] , \end{aligned} \qquad \text{(III.53a)}$$

and one can show that (e.g. (III.28b,c))

$$\mathcal{V}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] = \sum_{j=0}^{c-1} \sum_{\bar{l}_1 \cdots \bar{l}_K} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} \mathcal{V}\left[u_{i1,k}\right] + \left(j + \sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} u'_{i1,k}(1) \right)^2 \right\} p\left(j, \bar{l}_1, \cdots, \bar{l}_K\right) - \mathcal{M}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right]^2 .$$

$$\sum_{\{\bar{m}_1 \cdots \bar{m}_K\} \in \tilde{\Omega}} \frac{c(1-\rho)}{1 - E_{\bar{m}_1 \cdots \bar{m}_K}(1)} \mathcal{M}\left[\Psi_{\bar{m}_1 \cdots \bar{m}_K}\right] = c(1-\rho) \sum_{j=0}^{c-1} \sum_{\bar{l}_1 \cdots \bar{l}_K} \left(\sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} \sum_{n=2}^{L_k} \frac{u'_{in,k}(1)}{1 - \lambda_{n,k}(1)} \right) p\left(j, \bar{l}_1, \cdots, \bar{l}_K\right)$$

Finally, a rather straightforward calculation reveals that σ_s^2, the variance of the system content, is related to σ_q^2 via

$$\sigma_s^2 = \sigma_q^2 + \kappa_b \sigma_e^2 + 2c(1-\rho) \sum_{j=0}^{c-1} \sum_{\bar{l}_1 \cdots \bar{l}_K} \left(\sum_{k=1}^{K} \sum_{i=1}^{L} l_{i,k} u'_{i1,k}(1) \right) p\left(j, \bar{l}_1, \cdots, \bar{l}_K\right) \quad . \qquad \text{(III.53b)}$$

Evidently, one needs to invoke approximation (III.50) for the boundary probabilities in those traffic scenarios for which their exact calculation is unfeasible. If we confine ourselves to the single-server case c=1, then this produces the following approximations for the terms in the expressions for the mean and variance of the queue and system content that contain the boundary probabilities :

$$\begin{aligned}
\mathcal{M}\left[\Psi_{a,\bar{N}_1 \cdots \bar{N}_K}\right] &= \sum_{k=1}^{K} N_k \frac{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{u}'_{1,k}(1)}{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{I}_k} \\
\mathcal{V}\left[\Psi_{a,\bar{N}_1 \cdots \bar{N}_K}\right] &= \sum_{k=1}^{K} N_k \left\{ \frac{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \left(\bar{u}''_{1,k}(1) + \bar{u}'_{1,k}(1)\right)}{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{I}_k} - \left(\frac{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{u}'_{1,k}(1)}{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{I}_k} \right)^2 \right\}
\end{aligned} \qquad \text{(III.53c)}$$

$$\sum_{\{\bar{m}_1 \cdots \bar{m}_K\} \in \tilde{\Omega}} \frac{1}{1 - E_{\bar{m}_1 \cdots \bar{m}_K}(1)} \mathcal{M}\left[\Psi_{\bar{m}_1 \cdots \bar{m}_K}\right] \cong \sum_{k=1}^{K} N_k \sum_{n=2}^{L_k} \frac{\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{u}'_{n,k}(1)}{\left(1 - \lambda_{n,k}(1)\right)\left(\bar{\pi}_k^T \mathbf{Q}_k(0)^M \bar{I}_k\right)} \quad .$$

These approximate results can be relatively easy extended to the multi-server case and are well-suited for implementation in a computational algorithm; the formulae that are obtained in that case however are somewhat more tedious to write down, and will therefore not be explicitly recapitulated at this point.

A similar remark holds as far as the tail distribution of the queue and system content is concerned. The arguments of Section III.3.6.1 can be unabatedly reiterated, whereby the indices, which consist of the (homogeneous) set of integers $\bar{m}$, are replaced by the (heterogeneous) sets $\{\bar{m}_1 \cdots \bar{m}_K\}$. The poles that determine the tail behaviour of the queue and system content for instance will now be the solutions of

$$z_{0,\bar{m}_1 \cdots \bar{m}_K}^c = \prod_{k=1}^{K} \prod_{j=1}^{L_k} \lambda_{j,k}\left(z_{0,\bar{m}_1 \cdots \bar{m}_K}\right)^{m_{j,k}} \quad , \qquad \text{(III.54a)}$$

on the positive real axis, and larger than 1. Such a solution can be found in the region $]1,\mathcal{R}[$ for those values of $\{\bar{m}_1 \cdots \bar{m}_K\}$ for which the inequality

$$\lim_{z \to \mathcal{R}} E_{\bar{m}_1 \cdots \bar{m}_K}(z) \Big/ z^c > 1 \quad , \qquad \text{(III.54b)}$$

holds. The associated coefficients then satisfy (e.g. (A.13b))

$$\begin{aligned}\zeta_{\bar{m}_1\cdots\bar{m}_K} &\triangleq \lim_{z\to z_{o,\bar{m}_1\cdots\bar{m}_K}}\left(1-\frac{z}{z_{o,\bar{m}_1\cdots\bar{m}_K}}\right)Q(z)\\ &= -\frac{c(1-\rho)\left(z_{o,\bar{m}_1\cdots\bar{m}_K}-1\right)}{cz^{c}_{o,\bar{m}_1\cdots\bar{m}_K}-z_{o,\bar{m}_1\cdots\bar{m}_K}E'_{\bar{m}_1\cdots\bar{m}_K}\left(z_{o,\bar{m}_1\cdots\bar{m}_K}\right)}\Psi_{\bar{m}_1\cdots\bar{m}_K}\left(z_{o,\bar{m}_1\cdots\bar{m}_K}\right)\ .\end{aligned} \tag{III.54c}$$

The expressions (III.32c) that allow us to calculate an asymptotic approximation for the ccdf of the queue and system content basically remain unchanged,

$$\Pr[q>T_h]=\Pr[s>T_h+c]\cong\sum_{\bar{m}_1\cdots\bar{m}_K}\frac{\zeta_{\bar{m}_1\cdots\bar{m}_K}}{z_{o,\bar{m}_1\cdots\bar{m}_K}-1}z_{o,\bar{m}_1\cdots\bar{m}_K}^{-T_h}\ \text{, for large enough } T_h\text{ ,} \tag{III.54d}$$

where the sum over $\bar{m}_1\cdots\bar{m}_K$, in principle, runs over all sets for which (III.54b) is valid.

III.5.3 the packet sojourn time

From considerations such as the ones presented in Section III.4, it became apparent that the amount of time that a packet spends in the output buffer (under a FCFS queuing discipline) is determined by the queue content at the beginning of the packet's arrival slot, and the number of packets that have arrived during the same slot but are positioned before it in the queue. Consider a heterogeneous packet arrival process, whereby the order in which the packets that arrive during a slot are positioned in the queue is arbitrary, regardless of the traffic class they belong to (i.e., each packet has equal probability of being placed first, second, third, etc…) – this scenario will be referred to as *arbitrary order* (AO). However, in an ATM switching element, it can for instance be the case that the inlets are read into the output buffer in a fixed order during a slot. Hence, another credible scenario would be that all packets of type t are positioned before all packets of type t' in an output buffer, if $t< t'$, while all packet arrivals of the same type are placed in the queue in random order – this will be referred to as the *fixed-order-by-class* (FOC) scenario. Consequently, whatever the case, the sojourn time in the buffer of a packet of type t, $1\le t\le K$, will – if only slightly – depend on the value of t, due to the correlation and differentiation embedded in the traffic streams of different types.

Nevertheless, the waiting time and/or delay of a packet is most likely to be primarily determined by the amount of packets that were already in the system at the start of its arrival slot, and less so by the details of the ordering process of those packets that arrive during the same slot. Hence, for practical purposes, it will in all likelihood suffice to focus on the delay of an arbitrary packet, *regardless* of the traffic class it belongs to or the specifics of the ordering process, to deduce the main sojourn time performance indices. Moreover, if we consider an arbitrary packet under the AO ordering scheme, the results of Section III.4 can still be applied, implying that the waiting time / delay of an arbitrary packet can be unambiguously expressed in terms of the queue / system content at the beginning of an

arbitrary slot, and the relations between the moments and (tail) distribution of these drvs that were established in that section remain valid. Notwithstanding these remarks, the derivation of the class-dependent waiting time and/or delay characteristics for a heterogeneous packet arrival process with an AO and FOC ordering process, and solving the computational difficulties that arise accordingly, remains an interesting topic. In the course of this section, we will outline how these can be dealt with, without elaborating too much on all intricacies of the derivations. The purpose is to highlight the path that needs to be followed in order to obtain performance results for the type-t packet waiting time, rather than to carry out all the calculations in each and every detail (in particular with respect to its variance).

First, consider a packet of type t that enters the buffer under either the AO of FOC ordering discipline. Adopting similar notions as in Section III.4.1, with the addition of the subscript t to highlight the dependency on the traffic class t that the tagged packet belongs to, then from the relation

$$w_t = \left\lfloor \frac{r_t}{c} \right\rfloor \; ; \; r_t \triangleq q_t^* + f_t \quad ,$$

the following relations between the pmf and ccdf of the drvs w_t and r_t can be proved to hold

$$\begin{aligned} \Pr[w_t = i] &= \sum_{j=0}^{c-1} \Pr[r_t = ic + j] \\ \Pr[w_t > T_h] &= \Pr[r_t \geq (T_h + 1)c] \end{aligned} \quad , \tag{III.55a}$$

where the details of the ordering process and the dependence on t will be reflected by the pmf of r_t, and will be dealt with at the next stage of these derivations. By z-transforming the former relation, an equivalent relation as (II.42a) can be derived between the steady-state pgfs of w_t and r_t, i.e.,

$$W_t\left(z^c\right) = z^{-c} \sum_{n=0}^{c-1} U_c\left(z\eta^n\right) \mathbf{E}\left[\left(z\eta^n\right)^{q_t^* + f_t}\right] = z^{-c} \sum_{n=0}^{c-1} U_c\left(z\eta^n\right) R_t\left(z\eta^n\right) \quad .$$

Consequently, letting the $\mathcal{M}[\cdot]$ and $\mathcal{V}[\cdot]$ operators act upon this expression and applying similar calculation rules as in Section II.4.2, then the mean and variance of the type-t packet waiting time can be computed as

$$\mu_{w_t} = \frac{\mu_{r_t}}{c} - \frac{c-1}{2c} + \frac{1}{c} \sum_{n=1}^{c-1} \frac{\eta^n}{\eta^n - 1} R_t\left(\eta^n\right) \tag{III.55b}$$

$$\sigma_{w_t}^2 = \frac{\sigma_{r_t}^2}{c^2} + \frac{c^2-1}{12c^2} + \frac{1}{c^2}\sum_{n=1}^{c-1}\frac{\eta^n}{\eta^n-1}\left\{2\eta^n R_t'(\eta^n) - \left(2\mu_{r_t} + \frac{\eta^n+1}{\eta^n-1}\right)R_t(\eta^n)\right\} \qquad \text{(III.55c)}$$
$$-\left[\frac{1}{c}\sum_{n=1}^{c-1}\frac{\eta^n}{\eta^n-1}R_t(\eta^n)\right]^2 .$$

Note that these relations do not depend on the specific nature of the (heterogeneous) packet arrival process, whether it is a D-BMAP or some other type of correlated in packet arrival stream. In the single-server case c=1, these formulae reduce to trivial relations between the mean and variance of w_t and r_t due to $w_t \equiv r_t$ under these circumstances, but unfortunately for c>1 require the evaluation of $R_t(\cdot)$ and $R_t'(\cdot)$ in the points η^n, $1 \le n \le c$-1, that are positioned on the complex unit circle. This may be a task that is computationally demanding, for instance in case of a D-BMAP with a large state space. We would like to point out however that one can for instance prove the identity

$$\sum_{n=1}^{c-1}\frac{\eta^n}{\eta^n-1}R_t(\eta^n) = \frac{c-1}{2} - \sum_{i=0}^{\infty}\sum_{j=0}^{c-1} j \Pr[r_t = ic+j] \quad ,$$

which allows us to establish an upper- and lower bound for this term that occurs in the expression for the mean type-t packet waiting time, by replacing the factor j that appears in the sum in the right-hand side by 0 and c-1 respectively. Similar bounds, albeit less accurate, can be established as well for the sums that contain $R_t(\eta^n)$ and $R_t'(\eta^n)$ in the expression for the variance of the type-t waiting time, but are much harder to come by, and this topic will not be pursued here any further.

Equations (III.55a-c) make clear that the mean, variance and (tail) distribution of the class-t packet waiting time can be computed provided that we are able to derive expressions for the corresponding performance measures of the drv r_t, which, among others, will depend on the value of t and the ordering process under consideration, as explained before. In the remainder, when appropriate, we will add the subscript $_a$ or $_f$ to the random variables w_t, r_t, q_t^* and f_t (which, as before, represent the queue content at the beginning of an arbitrary type-t packet's arrival slot, and the number of packets that have arrived during the same slot and will be transmitted before it respectively) in order to explicitly refer to either the AO or FOC ordering process of packets that arrive during the same slot.

III.5.3.1 arbitrary order

Consider an arbitrary packet that belongs to the t-th traffic class and enters the buffer under the AO paradigm. The drv $f_{t,a}$, defined above, will then be a uniformly distributed drv that is

bounded by 0 and the total number of packet arrivals minus 1, regardless of the queue content at the start of the tagged packet's arrival slot. Hence, we may write

$$\mathbf{E}\left[x^{f_{t,a}} \middle| q_{t,a}^* = j, \sum_{k=1}^{K} e_k^* = i, e_t^* = i_t \right] = \frac{x^i - 1}{i(x-1)} \quad , \quad i \geq i_t > 0 \quad .$$

On the other hand, since the tagged packet has been arbitrarily picked among all type-t packet arrivals, we can use similar arguments as in Section II.4.1 to show that

$$\Pr\left[q_{t,a}^* = j, \sum_{k=1}^{K} e_k^* = i, e_t^* = i_t \right] = \frac{i_t}{c\rho_t} \Pr\left[q = j, \sum_{k=1}^{K} e_k = i, e_t = i_t \right] \quad ,$$

where $c\rho_t$ represents the average number of type-t packet arrivals per slot. Hence, if we denote by the drv e the total number of packet arrivals during an arbitrary slot, then the combination of the latter two expressions results in the following expression for the pgf of $r_{t,a}$

$$R_{t,a}(z) = \mathbf{E}\left[z^{q_{t,a}^* + f_{t,a}} \right] = \frac{1}{c\rho_t} \mathbf{E}\left[e_t \cdot z^q \cdot \frac{z^e - 1}{e(z-1)} \right] \quad .$$

If we define $R_a(z)$ as the pgf of r_a, the number of packets in the queue at the arrival epoch of an *arbitrary* packet (i.e., regardless of its type t) under the AO scheme, we may then write

$$R_a(z) = \frac{1}{c\rho} \sum_{t=1}^{K} c\rho_t R_{t,a}(z) = \frac{1}{c\rho} \mathbf{E}\left[\sum_{t=1}^{K} e_t \cdot z^q \cdot \frac{z^e - 1}{e(z-1)} \right] = \frac{1}{c\rho(z-1)} \mathbf{E}\left[z^q \cdot \left(z^e - 1 \right) \right] \quad ,$$

and a comparison with (III.33), combined with the calculations that follow thereafter, reveals that the waiti arrivals case, implying that the results that were derived in Sections III.4.1-3 can still be applied, as indicated before.

By taking the appropriate derivatives with respect to z for z=1 of the penultimate identity, we can establish the following expressions for the mean and variance of $r_{t,a}$

$$\begin{aligned} c\rho_t \cdot \mu_{r_{t,a}} &= \mathbf{E}[q \cdot e_t] + \left(\sigma_{e_t}^2 + c\rho_t (c\rho - 1) \right) \Big/ 2 \\ c\rho_t \cdot \sigma_{r_{t,a}}^2 &= \mathbf{E}\left[q^2 \cdot e_t \right] + \mathbf{E}[q \cdot e_t (e-1)] + \mathbf{E}[e_t (e-1)(2e-1)]/6 - c\rho_t \cdot \mu_{r_{t,a}}^2 \quad . \end{aligned} \tag{III.56a}$$

The latter two expressions for the moments of $r_{t,a}$ are generally valid, in the sense that they do not depend on the correlation structure of the packet arrival process.

In order to derive expressions for the expected values that occur in the right-hand sides of these formulae, we need to establish an expression for the joint pgf

$$H_{t,a}(z,x,y) \triangleq \mathbf{E}\left[z^q x^e y^{e_t} \right] \quad .$$

Observe that the set of drvs (q,e,e_t) relate to the queue content at the beginning of an arbitrary slot and the number of arrivals during *the same* slot, while the system's state description, described by the joint pgf $P_q(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)$, pertains to the queue content at the start of an arbitrary slot, and the state of the Markov chain that defines the packet arrival during the *preceding* slot. Hence, we can invoke identity (III.41a) to write

$$H_{t,a}(z,x,y)=P_q\left(z,\mathbf{Q}_1(x)\bar{\boldsymbol{I}}_1,\cdots,\mathbf{Q}_{t-1}(x)\bar{\boldsymbol{I}}_{t-1},\mathbf{Q}_t(xy)\bar{\boldsymbol{I}}_t,\mathbf{Q}_{t+1}(x)\bar{\boldsymbol{I}}_{t+1},\cdots,\mathbf{Q}_K(x)\bar{\boldsymbol{I}}_K\right) \quad ,$$

which, in view of (III.48c), becomes

$$H_{t,a}(z,x,y)=c(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)\Theta_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,t,a}(z,x,y)}{z^c-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)$$

$$\Theta_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,t,a}(z,x,y)\triangleq\left(\mathbf{W}_t(z)\mathbf{Q}_t(xy)\bar{\boldsymbol{I}}_t\right)^{\bar{\boldsymbol{m}}_t}\prod_{\substack{k=1\\k\neq t}}^{K}\left(\mathbf{W}_k(z)\mathbf{Q}_k(x)\bar{\boldsymbol{I}}_k\right)^{\bar{\boldsymbol{m}}_k} \quad .$$

Also note that the joint pgf of e and e_t is given by $H_{t,a}(1,x,y)$, and satisfies

$$H_{t,a}(1,x,y)=\left(\bar{\boldsymbol{\pi}}_t^T\mathbf{Q}_t(xy)\bar{\boldsymbol{I}}_t\right)^{N_t}\prod_{\substack{k=1\\k\neq t}}^{K}\left(\bar{\boldsymbol{\pi}}_t^T\mathbf{Q}_k(x)\bar{\boldsymbol{I}}_k\right)^{N_k} \quad .$$

The mixed expected values that contain the drv q occurring in the right-hand side of (III.56a), can now be explicitly calculated by taking the appropriate derivatives of $H_{t,a}(z,x,y)$. For instance, if we take the derivative with respect to z and y for $z=x=y=1$, and keep in mind the computational guidelines that were mentioned in Section III.5.2.2 for calculating these kind of derivatives, we obtain the following expression for $\mathbf{E}[q\cdot e_t]$

$$\begin{aligned}\mathbf{E}[q\cdot e_t]&=\mu_q\cdot\frac{\partial}{\partial y}\left(\bar{\boldsymbol{\pi}}_t^T\mathbf{Q}_t(y)\bar{\boldsymbol{I}}_t\right)^{N_t}\Big|_{y=1}+\frac{\partial^2}{\partial y\partial z}\left(\bar{\boldsymbol{w}}_{1,t}(z)\mathbf{Q}_t(y)\bar{\boldsymbol{I}}_t\right)^{N_t}\Big|_{z=y=1}+\chi_t\\&=c\rho_t\cdot\mu_q+N_t\bar{\boldsymbol{w}}'_{1,t}(1)\mathbf{Q}'_t(1)\bar{\boldsymbol{I}}_t+\chi_t\\&=c\rho_t\cdot\mu_q+\frac{(\kappa_{b,t}-1)\sigma_{e_t}^2}{2}+\chi_t \quad ,\end{aligned}$$

where we have also taken into account the properties $\bar{\boldsymbol{w}}'_{1,t}(1)\bar{\boldsymbol{I}}_t=\bar{\boldsymbol{w}}''_{1,t}(1)\bar{\boldsymbol{I}}_t=0$ (due to $\bar{\boldsymbol{w}}_{1,t}(z)\bar{\boldsymbol{I}}_t=1$) in the evaluation of the second-order derivative with respect to z for $z=1$ of the relation $\bar{\boldsymbol{w}}_{1,t}(z)\mathbf{Q}_t(z)\bar{\boldsymbol{I}}_t=\lambda_{1,t}(z)\bar{\boldsymbol{w}}_{1,t}(z)\bar{\boldsymbol{I}}_t$. The contribution χ_t stems from the derivative – with respect to z and y for $z=y=1$ – of the terms in the sum in the right-hand side of the expression for $H_{t,a}(z,1,y)$, for which the first component of $\bar{\boldsymbol{m}}_t$ is equal to N_t-1. Deriving

closed-form expression for χ_t requires a punctual evaluation of these derivatives. By making use (among others) of relations such as

$$\begin{aligned} &\frac{d}{dz}\left(u_{ij,k}(z)\overline{\boldsymbol{w}}_{j,k}(z)\overline{\boldsymbol{I}}\right)\Big|_{z=1} = u'_{ij,k}(1) \\ &\left(u_{ij,k}(1)\overline{\boldsymbol{w}}_{j,k}(1)\right)\mathbf{Q}'_k(1)\overline{\boldsymbol{I}} = -\left(1-\lambda_{j,k}(1)\right)u'_{ij,k}(1) \end{aligned} \quad , \quad 2 \le j \le L_k \quad ,$$

which immediately follow from taking the appropriate derivatives of equations such as (III.7e), then one is able to establish after some tedious calculations that

$$\chi_t = c(1-\rho) \sum_{\overline{\boldsymbol{I}}_1 \cdots \overline{\boldsymbol{I}}_K} \sum_{i=1}^{L_t} \left\{ l_{i,t} u'_{i1,t}(1) \right\} \sum_{j=0}^{c-1} p\left(j, \overline{\boldsymbol{I}}_1, \cdots, \overline{\boldsymbol{I}}_K\right) \quad .$$

Consequently, the latter expression for $\mathbf{E}[q \cdot e_t]$ and (III.56a) yield the following result for the mean value of $r_{t,a}$:

$$\mu_{r_{t,a}} = \mu_q + \frac{c\rho}{2} + \frac{\kappa_{b,t}\sigma_{e_t}^2}{2c\rho_t} - \frac{1}{2} + \frac{1}{c\rho_t}\chi_t \quad . \tag{III.56b}$$

An expression for the variance of $r_{t,a}$ can be derived in the same manner, but will not be pursued here. Let us instead, as a check on our calculations, focus on the single-server case and derive an expression for μ_w, the mean waiting time of an *arbitrary* packet, regardless of the class it belongs to. In view of the AO ordering mechanism, this quantity should satisfy Little's result, i.e., $\rho\mu_w = \mu_q$. Considering $c=1$, then from (III.55a) and (III.56b), and the latter expression for χ_t, we deduce that

$$\mu_w = \sum_{t=1}^{K} \frac{\rho_t}{\rho} \mu_{w_{t,a}} = \sum_{t=1}^{K} \frac{\rho_t}{\rho} \mu_{r_{t,a}} = \mu_q + \frac{\rho}{2} + \frac{\kappa_b \sigma_e^2}{2\rho} - \frac{1}{2} + \frac{1-\rho}{\rho} \boldsymbol{\mathcal{M}}\left[\Psi_{\overline{\boldsymbol{N}}_1 \cdots \overline{\boldsymbol{N}}_K}\right] \quad ,$$

where we have also taken equation (III.44a) for the burst factor κ_b and expression (III.52a) for $\boldsymbol{\mathcal{M}}\left[\Psi_{\overline{\boldsymbol{N}}_1 \cdots \overline{\boldsymbol{N}}_K}\right]$ (with $c=1$) into consideration. Since the mean queue content in the single-server case equals

$$\mu_q = \frac{\kappa_b \sigma_e^2}{2(1-\rho)} - \frac{\rho}{2} + \boldsymbol{\mathcal{M}}\left[\Psi_{\overline{\boldsymbol{N}}_1 \cdots \overline{\boldsymbol{N}}_K}\right] \quad ,$$

it is easily verified that the latter two equations indeed imply that $\rho\mu_w = \mu_q$.

As far as the calculation of the tail distribution of $r_{t,a}$ is concerned, let us commence by pointing out that $R_{t,a}(z)$ can be calculated as

$$R_{t,a}(z)=\frac{1}{c\rho_t(z-1)}\frac{\partial}{\partial y}\sum_{x=1}^{z}x^{-1}H_{t,a}(z,x,y)dx\Big|_{y=1}=\sum_{\bar{m}_1\cdots\bar{m}_K}\frac{c(1-\rho)(z-1)\tilde{\Theta}_{\bar{m}_1\cdots\bar{m}_K,t,a}(z)}{z^c-E_{\bar{m}_1\cdots\bar{m}_K}(z)}\Psi_{\bar{m}_1\cdots\bar{m}_K}(z)$$

$$\tilde{\Theta}_{\bar{m}_1\cdots\bar{m}_K,t,a}(z)\triangleq\frac{1}{c\rho_t(z-1)}\sum_{x=1}^{z}\left(\sum_{i=1}^{L_t}m_{i,t}\frac{\bar{w}_{i,t}(z)\mathbf{Q}'_t(x)\bar{I}_t}{\bar{w}_{i,t}(z)\mathbf{Q}_t(x)\bar{I}_t}\prod_{k=1}^{K}\prod_{j=1}^{L_k}\left(\bar{w}_{j,k}(z)\mathbf{Q}_k(x)\bar{I}_k\right)^{m_{j,k}}\right)dx\ .$$

Adopting the multiple-pole approximation procedure for calculating the ccdf tail asymptote of $r_{t,a}$, that has become the standard approach in this chapter, we thus obtain

$$\Pr\left[r_{t,a}>T_h\right]\cong\sum_{\bar{m}_1\cdots\bar{m}_K}\frac{\tilde{\Theta}_{\bar{m}_1\cdots\bar{m}_K,t,a}\left(z_{o,\bar{m}_1\cdots\bar{m}_K}\right)\zeta_{\bar{m}_1\cdots\bar{m}_K}}{z_{o,\bar{m}_1\cdots\bar{m}_K}-1}z_{o,\bar{m}_1\cdots\bar{m}_K}^{-T_h}\quad,\qquad\text{(III.56c)}$$

where the dominant poles $z_{o,\bar{m}_1\cdots\bar{m}_K}$ and the associated coefficients can be calculated as outlined in Section III.5.2.2. If, for each traffic class k, the D-BMAP under consideration is a MMBP, then the integrand in the expression for $\tilde{\Theta}_{\bar{m}_1\cdots\bar{m}_K,t,a}(\cdot)$ is a polynomial in the variable x of degree $\sum_k L_k-1$, and the integration can, in principle, be carried out explicitly. For a general D-BMAP however, the integral cannot be worked out as a closed-form formula, and must be computed numerically.

Summarising, if we combine the equations (III.55a-c) with (III.56a-c), then this allows us, at least in principle, to calculate the mean, variance, and the asymptotic tail behaviour for the ccdf of the waiting time of an arbitrary packet of type t in case of an AO ordering procedure for those packets that arrive during the same slot as a tagged one. However, if we want to calculate the per-class waiting time performance characteristics, a couple of difficulties arise that may inhibit the practical use of the formulae presented in this subsection. The first type is inherent to the multi-server nature of the model under consideration, and stems from the need to compute the pgf $R_{t,a}(\cdot)$ and its derivative $R'_{t,a}(\cdot)$ in the arguments η^n, $1\le n\le c$-1, in the evaluation of the mean and variance of the type-t waiting time. Another impediment that must be overcome is induced by the, possibly numerical, evaluation of the integral in the expression for $\tilde{\Theta}_{\bar{m}_1\cdots\bar{m}_K,t,a}(\cdot)$ when we calculate the tail distribution of the type-t waiting time ccdf. This is specific for the AO ordering mechanism that we have focussed on up to now, and can be alleviated by considering alternative ordering schemes, such as the one treated next.

III.5.3.2 fixed-order-by-class

We now consider an arbitrary packet of type t that enters the buffer during a slot under the FOC ordering scheme, meaning that all packets of type t', $t'<t$, that arrive during the same slot will be positioned in the queue before all type-t packet arrivals, while all packets of type t'',

$t<t''$, that arrive during the course of this slot will be placed in the queue behind the type-t packets. Among the type-t packets that enter the system during the tagged packet's arrival slot, an arbitrary ordering scheme remains valid. This implies that $f_{t,f}$, the drv that captures all packets that have arrived during the same slot as the tagged one and will be positioned before it in the queue, consists of the sum of all type t' packet arrivals with $t'<t$, and a second component that is a uniformly distributed drv which is bounded by 0 and the total number of type-t packet arrivals minus 1. Hence, for this ordering scenario we may write

$$\mathbf{E}\left[x^{f_{t,f}} \middle| q^*_{t,f} = j, \sum_{k=1}^{t-1} e^*_k = i, e^*_t = i_t \right] = x^i \frac{x^{i_t}-1}{i_t(x-1)} \quad ; \quad i \geq 0 \; ; \; i_t > 0 \quad .$$

On the other hand, it remains so that the tagged packet has been arbitrarily picked among all type-t packet arrivals, meaning that the relation

$$\Pr\left[q^*_{t,f} = j, \sum_{k=1}^{t-1} e^*_k = i, e^*_t = i_t \right] = \frac{i_t}{c\rho_t} \Pr\left[q = j, \sum_{k=1}^{t-1} e_k = i, e_t = i_t \right] \quad ,$$

holds. Therefore, if we denote by the drv $\tilde{e}_t$, with mean $c\tilde{\rho}_t$, the sum of all type-t' arrivals, $t'<t$, during the tagged packet's arrival slot, then the combination these two identities results in the following expression for the steady-state pgf $R_{t,f}(z)$ of $r_{t,f}$

$$R_{t,f}(z) = \mathbf{E}\left[z^{q^*_{t,f} + f_{t,f}} \right] = \frac{1}{c\rho_t} \mathbf{E}\left[z^{q+\tilde{e}_t} \cdot \frac{z^{e_t}-1}{(z-1)} \right] \quad .$$

By taking the appropriate derivatives with respect to z for $z=1$ of both hand sides of the above equation, we can deduce that the mean and variance of $r_{t,f}$ satisfy

$$\begin{aligned} c\rho_t \cdot \mu_{r_{t,f}} &= \mathbf{E}[q \cdot e_t] + c^2 \rho_t \tilde{\rho}_t + \left(\sigma^2_{e_t} + c\rho_t (c\rho_t - 1) \right) \Big/ 2 \\ c\rho_t \cdot \sigma^2_{r_{t,f}} &= \mathbf{E}\left[(q+\tilde{e}_t)^2 \cdot e_t \right] + \mathbf{E}[(q+\tilde{e}_t) \cdot e_t (e_t - 1)] + \mathbf{E}[e_t (e_t - 1)(2e_t - 1)] \big/ 6 - c\rho_t \cdot \mu^2_{r_{t,f}} \quad . \end{aligned} \tag{III.57a}$$

We would like to point out once more that the latter two expressions for the moments of $r_{t,f}$ do not depend on the specifics of the (correlated) packet arrival process. Naturally, the expression that was derived before for $\mathbf{E}[q \cdot e_t]$ remains valid, which results in the following expression for the mean value of $r_{t,f}$ under the FOC paradigm

$$\mu_{r_{t,f}} = \mu_q + \frac{\kappa_{b,t} \sigma^2_{e_t}}{2c\rho_t} + c\tilde{\rho}_t + (c\rho_t - 1)/2 + \frac{1}{c\rho_t} \chi_t \quad . \tag{III.57b}$$

Observe that this implies that

$$\mu_{r_{t,f}} - \mu_{r_{t,a}} = c(\tilde{\rho}_t - (\rho - \rho_t)/2) = \frac{c}{2}\left(\sum_{k=1}^{t-1} \rho_k - \sum_{k=t+1}^{K} \rho_k \right) \quad .$$

Hence, since the type-t packet waiting time satisfies $w_t = \lfloor r_t / c \rfloor$ for both paradigms, the difference between the respective mean packet waiting times will typically be no more than 0.5 slots.

In order to calculate the remaining expected values that occur in the right-hand side of the expression (III.57a) for $\sigma^2_{r_{t,f}}$, we need to establish an expression for the joint pgf

$$H_{t,f}(z,y) \triangleq \mathbf{E}\left[z^{q+\tilde{e}_t} y^{e_t} \right] \quad .$$

Adopting a similar calculus as in the AO-case, we can show that

$$H_{t,f}(z,y) = P_q\left(z, \mathbf{Q}_1(z)\overline{\boldsymbol{I}}_1, \cdots, \mathbf{Q}_{t-1}(z)\overline{\boldsymbol{I}}_{t-1}, \mathbf{Q}_t(y)\overline{\boldsymbol{I}}_t, \overline{\boldsymbol{I}}_{t+1}, \cdots, \overline{\boldsymbol{I}}_K\right) \quad ,$$

which, in view of (III.48c), can be transformed into

$$H_{t,f}(z,y) = c(1-\rho) \sum_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K} \frac{(z-1)\Theta_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K, t, f}(z,y)}{z^c - E_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K}(z)} \Psi_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K}(z)$$

$$\Theta_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K, t, f}(z,y) \triangleq \left(\mathbf{W}_t(z)\mathbf{Q}_t(y)\overline{\boldsymbol{I}}_t\right)^{\overline{\boldsymbol{m}}_t} \prod_{k=1}^{t-1} \left(\mathbf{W}_k(z)\mathbf{Q}_k(z)\overline{\boldsymbol{I}}_k\right)^{\overline{\boldsymbol{m}}_k} \quad .$$

The remaining unknown quantities that occur in the right-hand side of expression (III.57a) for the variance of $r_{t,f}$ can then be computed by taking the appropriate derivatives of $H_{t,f}(z,y)$ with respect to z and y, for $z=y=1$.

For the purpose of establishing a multiple-pole approximation for the ccdf of $r_{t,f}$, note that $R_{t,f}(z)$ can be calculated from

$$R_{t,f}(z) = \frac{H_{t,f}(z,z) - H_{t,f}(z,1)}{c\rho_t(z-1)} = \sum_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K} \frac{c(1-\rho)(z-1)\tilde{\Theta}_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K, t, f}(z)}{z^c - E_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K}(z)} \Psi_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K}(z) \quad ,$$

$$\tilde{\Theta}_{\overline{\boldsymbol{m}}_1 \cdots \overline{\boldsymbol{m}}_K, t, f}(z) \triangleq \frac{1}{c\rho_t(z-1)} \left(\prod_{i=1}^{L_t} \lambda_{i,t}(z)^{m_{i,t}} - 1 \right) \left(\prod_{k=1}^{t-1} \prod_{j=1}^{L_k} \lambda_{j,k}(z)^{m_{j,k}} \right) ,$$

where we have exploited the property that $\mathbf{W}_k(z)\mathbf{Q}_k(z)\overline{\boldsymbol{I}}_k = \boldsymbol{\Lambda}_k(z)\mathbf{W}_k(z)\overline{\boldsymbol{I}}_k = \boldsymbol{\Lambda}_k(z)\overline{\boldsymbol{I}}_k$. The multiple-pole approximation for calculating the ccdf of $r_{t,f}$ thus yields

$$\Pr\left[r_{t,f} > T_h\right] \cong \sum_{\bar{m}_1 \cdots \bar{m}_K} \frac{\tilde{\Theta}_{\bar{m}_1 \cdots \bar{m}_K, t, f}\left(z_{0,\bar{m}_1 \cdots \bar{m}_K}\right) \zeta_{\bar{m}_1 \cdots \bar{m}_K}}{z_{0,\bar{m}_1 \cdots \bar{m}_K} - 1} z_{0,\bar{m}_1 \cdots \bar{m}_K}^{-T_h} \quad . \tag{III.57c}$$

Contrary to the AO ordering scenario, the coefficients $\tilde{\Theta}_{\bar{m}_1 \cdots \bar{m}_K, t, f}(\cdot)$ that occur in the right-hand side of this expression can be explicitly calculated without great difficulty.

The conclusion of this section may be a good opportunity to briefly comment on an ambiguous reasoning that is sometimes encountered in discussions or comments concerning the packet waiting time and/or delay in buffers with this type of heterogeneous arrival patterns. For the FOC ordering scheme, one could be inclined to intuitively accept that type t' packets typically experience smaller sojourn times in the buffer compared to type t packets if $t'<t$, since type t' packets have 'priority' over type t packets that arrive during the same slot. This, however, is not necessarily true. This can be explained as follows : consider two types of traffic generated by 2-state IBP (i.e., ON/OFF) traffic sources. Type-1 traffic is relatively bursty, which corresponds to ON and OFF periods that are typically long, while type-2 traffic is rather smooth. Under the FOC ordering scheme, type-1 packet arrivals are up for transmission before the type-2 packets that arrive during the same slot. Nonetheless, provided that there is a type-1 packet arrival, there are bound to have been a comparatively large number of type 1 packet arrivals during the preceding slots, due to the long ON periods for this specific traffic class. In other words, there is a stronger positive correlation between an arriving type-1 packet and the amount of work in the buffer at the beginning of its arrival slot compared to the type-2 packet arrivals, which may all but undo the advantage of having FOC priority. Focussing on the single-server case c=1, then from (III.57b) it immediately follows that the mean waiting time of type-1 packets will exceed the mean type-2 waiting time if

$$\frac{\kappa_{b,1}\sigma_{e_1}^2}{2\rho_1} + \frac{\chi_1}{\rho_1} > \frac{\kappa_{b,2}\sigma_{e_2}^2}{2\rho_2} + \frac{\chi_2}{\rho_2} + \frac{\rho}{2} \quad ,$$

and it is not too difficult to construct a traffic scenario where such is the case, by picking high enough values of $\kappa_{b,1}$ and ρ_2, and low enough values of $\kappa_{b,2}$ and ρ_1. In the multi-server case, the relation between the mean waiting time of the two traffic types is somewhat more muddled (due to the presence of $R_{t,f}(\cdot)$ in expression (III.55a) for the mean waiting time, that needs to be evaluated at η^n, $0 \le n \le c$-1, for the two traffic types), but the (qualitative) comments made above remain valid.

III.6 The tolerable load of a set of packet-based phones : a case study

In order to demonstrate the efficacy and applicability of the results for a buffer with a D-BMAP that have been deduced up to now, in the performance assessment of specific

components of communication networks, let us consider a setup that was reported in [99]. In this section, we will consider a scenario of packet-based phones with and without silence suppression, and the purpose is to study the impact of parameters such as the link rate (i.e., the available bandwidth) R_{link}, the activity grade α (i.e., the fraction of time that a phone is active), and the codec bit rate R_{cod}, on the maximum number of phones that can be supported by a network access node that is fed by such traffic.

Packetised voice transport (e.g., Voice-over-IP (VoIP), Voice-and-Telephony-Over-ATM (VTOA)) has several advantages over circuit-switched voice transport. The former is more flexible than the latter, because an active phone does not occupy a trunk (i.e., a fixed portion of the capacity) for the entire duration of a phone call, but capacity is dynamically shared with all other active phones (and possibly with data sources). Unlike the (circuit-switched) Public Switched Telephone Network (PSTN) which handles $64 kb/s$ voice streams encoded in the G.711 format, a packet-based voice network can carry any codec format that both communicating phones support. This is particularly an advantage, because the codec technology is evolving to and aiming at ever-smaller bit rates. Moreover, in contrast to on a circuit-switched network, silence suppression can be exploited on a packet-based network, reducing the average codec bit rate even further. However, a voice call routed over a packet-based network risks experiencing a degraded quality, because more delay and distortion are likely to be introduced in comparison to a call switched over the PSTN. The acceptable bounds on the delay and the distortion are known and reported in ([151], [187]). The remaining issue is how to dimension the network elements (the packetiser, the network nodes, the dejittering mechanism, etc.) such that these bounds are met ([150], [255]). This analysis concentrates on the delay introduced in one (access) node in the network.

The problem tackled here has some similarities with the dimensioning of a PSTN switch. Consider (circuit-switched) phones (that generate voice streams in the G.711 format) with an activity grade α. The activity grade is determined by the average call duration and the average passive period. Both the active and passive period are traditionally assumed to be exponentially distributed. Each active phone requires a trunk in the switch. If no trunk is available when a phone attempts a call, the call is blocked. Consequently, the (inverse) Engset formula ([226]), which stems from the continuous-time $M/M/N_{PSTN}/N_{PSTN}/N$ queuing model, calculates the number N of phones that can be supported by a switch with N_{PSTN} trunks (i.e., of capacity $R_{link}=N_{PSTN}\cdot 64 kb/s$) given a certain tolerated blocking probability. In addition, the (inverse) Erlang-B formula ([195]) can be used instead of the (inverse) Engset formula, if the number N of phones is relatively high compared to the number of trunks N_{PSTN}. In the current scenario however, we also consider phones with an activity grade α, which are now packet-based (hence, a discrete-time setting, with geometrically-distributed active and passive periods) and not necessarily encoded in the G.711 format. Calls are (in principle)

never blocked, but the packet sojourn times in the node should be restricted, which limits the number of phones that can be supported. Therefore, the number N of phones that can be supported by one packet-based multiplexing node of capacity R_{link} will be calculated, given that a appropriately chosen quantile of the packet waiting time is bounded by a certain preset threshold, and the methodology that has been developed in this dissertation for calculating these quantiles will be adopted. Naturally, although we concentrate on the converse problem (i.e., determine N given R_{link}) here, the same methodology developed here can be used as well to determine the minimal capacity R_{link} to support a given number N of phones.

It might be argued that because of the emergence of terabit networks, capacity will be for free, i.e., that future nodes will have so much capacity that they can support much more voice calls than they actually will have to. Although this might be the case for backbone networks, for access networks, resources will remain very much limited in the near foreseeable future. Even for backbone nodes that will transport voice and data, voice traffic will be transported in (one of) the high priority classes and it is likely that not more bandwidth than what is strictly necessary will be reserved for voice traffic.

III.6.1 packetised voice transport

In the packetised transport of digital voice there are three essential stages, as depicted in Fig. III-19. In the current scenario we focus on VoIP as an example, but it can be easily adapted for any type of packet-based network, albeit with different amounts of overhead per voice payload. In the first stage, the digital voice signal (i.e., voice that is sampled at e.g. 8 kHz and quantized with a uniform e.g. 13-bit quantiser) is encoded and packetised. We consider four codec bit rates R_{cod}=64kb/s, 32kb/s, 16kb/s and 8kb/s. In order to be able to easily compare the multiplexing behaviour of these different codecs, we settle for a payload size of 160B(ytes). This means that the packetisation delay (i.e., the time that is required to fill 160B) will equal 20ms, 40ms, 80ms and 160ms, respectively. A packetisation delay of 20ms for the 64kb/s is quite reasonable, but a packetisation delay of 80ms and 160ms for the 16kb/s and 8kb/s codec respectively is a bit on the high side. However, it is not our aim at this point to determine the optimal packetisation delay. Because in VoIP the header consists of 20 IP B(ytes), 8 UDP B and 12 RTP B, the size of all voice packets is equal to 200B. As a result, when a phone is active (and talking), it produces a flow of IP packets of 200B with a inter-packet time that is equal to 20ms, 40ms, 80ms and 160ms, for the 64kb/s, 32kb/s, 16kb/s and 8kb/s codec, respectively.

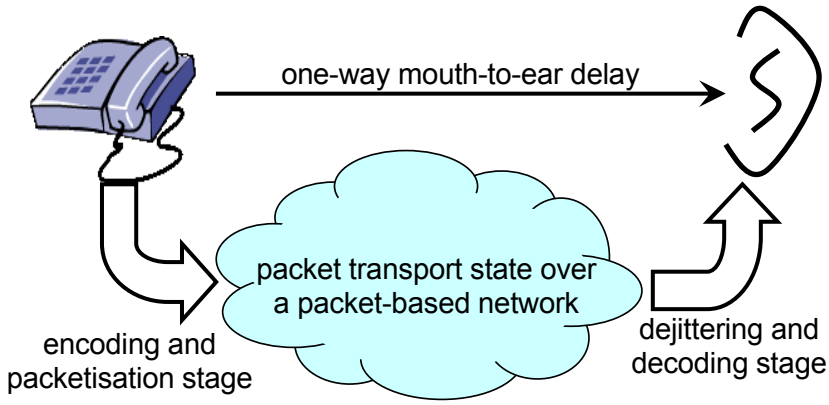


Figure III-19 : stages in the packetised transport of voice

In the second stage, this flow of packets is transported over an IP network consisting of several access and backbone nodes. In the transport of the voice flow over this network some delay is incurred. The network delay can be split into two parts : a deterministic part, referred to as the minimal network delay, and a stochastic part, referred to as the (total) packet waiting time. The minimal network delay mainly consists of the propagation delay (of $5\mu s$ per km), the sum of all serialization delays, the route look-up delay, etc. The total packet waiting time is the sum of the waiting times that a packet encounters in each node. The waiting time in one network node stems from the competition of the packets of several flows for the available resources in that particular node.

Our aim is to study the waiting time that the voice packets incur in one of the network nodes. We consider link rates R_{link} from $512 kb/s$ up to $10.24 Mb/s$, which are typical link rates in an access network. Moreover, as will become clear from the results below, at a link rate of $10.24 Mb/s$, the tolerable load reaches some sort of limiting behaviour. If there is only one bottleneck node on the mouth-to-ear path, the waiting time incurred in this node practically solely determines the total waiting time. If there are more than one nodes with a significant contribution to the total queuing delay, the individual sojourn time statistics need to be combined, e.g., by a convolution approach that originates from assuming that the packet sojourn times in consecutive nodes are (close to being) statistically independent. Nevertheless, such a scenario falls beyond the scope of this brief example.

In the final stage, the distorted packet flow is dejittered and decoded. Since the decoder needs the packets at a constant rate, dejittering is absolutely necessary. Dejittering a voice flow consists of retaining the fastest packets in the dejittering buffer to allow the slowest ones to catch up, and the fastest packets are the ones that do not have to wait for transmission in any of the nodes. So, in principle, the fastest packets have to be retained in the dejittering buffer for a time equal to the maximal total waiting time. Because voice codecs can tolerate some packet loss and because waiting for the slowest packet frequently introduces too much delay, it is often so that the fastest packets are retained in the dejittering buffer for a time

equal to the 10^{-X}-quantile of the total waiting time, as defined by (II.34). This means that a fraction 10^{-X} of the packets will be lost, because they arrive too late. It depends on the codec to what extent packet loss can be tolerated. Typical values for 10^{-X} lie in the interval $[10^{-5},10^{-2}]$ (see [151]). Here, we take a value of X=4. For more details on the dejittering paradigm, we refer to [42], [115], [240] and the references therein.

From the above explanation it is clear that the waiting time incurred in a network node contributes to the mouth-to-ear delay. This example focuses on the waiting time introduced in a (bottleneck) access node. More specifically, we determine up to which value of the load ρ the node can be loaded for the 10^{-X}-quantile of the packet waiting time to reach a specific value T_d. In view of the previous, this also implies that if the voice packet stream is delayed in the dejittering buffer over an amount of time that is equal to T_d, then no more than a fraction of 10^{-X} packets will arrive too late for a timely playout. In most cases we use T_d=12.5ms. The choice for this value is reasonable, although somewhat conservative. It is well known that if the echo is perfectly controlled, the bound on the mouth-to-ear delay to ensure an interactive call is 150ms, although this bound is not strict and may be exceeded under certain circumstances ([151],[187]). If we subtract the main components (i.e., the codec delay, the packetisation delay, the sum of all serialization delays, the propagation delay, etc.) from this available budget of 150ms, and take into account that additional network nodes might also (slightly) contribute to the total packet waiting time, we end up with a waiting time budget for the access node in the neighbourhood of the value of T_d mentioned above.

III.6.2 the queuing model

III.6.2.1 modelling the phones

In order to assess the delay incurred in one network node, we study the following discrete-time queuing model. The time unit (i.e. slot) is taken equal to the time to insert a voice packet on the link. Hence, the transmission time for each packet is 1 slot. Remark that the time unit depends on the link rate R_{link} and the packet size (always equal to 200B here), and hence, ranges between 156.25μs and 3.125ms for R_{link} 10.24Mb/s and 512kb/s, respectively. The buffer in the network node has one channel that transmits packets at a rate of R_{link}, and the buffer size is assumed to be sufficiently large to accommodate the incoming packets, which amounts to assuming that packet loss due to buffer overflow is small compared to the loss of information due to packets that arrive too late in the dejittering buffer at the receiver side.

We will model the traffic that is generated by a phone as a MMBP, where we let the states of the MMBP correspond with the possible modi of an IP-phone conversation. If the phone is passive (or silent), it sends no packets at all. When the phone is in the active (and talking) state, it behaves as a Bernoulli source, i.e., it generates a packet with probability $1/I_a$ during each slot, with I_a the (mean) inter-packet time (expressed in slots),

$$I_a = \phi \frac{R_{link}}{R_{cod}} \quad , \tag{III.58}$$

where ϕ represents the filling factor of the packets, which signifies the payload size divided by the packet size. In the current scenario, this parameter is equal to 0.8 (=160/200). In the upcoming analysis, we consider phones that do not, and phones that do, use silence suppression respectively.

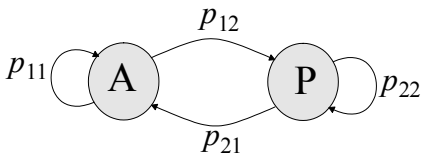


Figure III-20 : state transition diagram for a phone without silence suppression.

a phone without silence suppression

In this scenario, a phone can be either in an active or passive state. We assume that a phone is in the active state 'A' with probability α (and in the passive state 'P' with probability 1-α) and that the average duration of a call is 2 minutes (120*s*). A benchmark value for α is 0.1, which corresponds to a call attempt rate of 3 calls per hour. In the following, quantities such as the average call duration T_A and the average passive period T_P are expressed in slots, and the conversion from seconds to slots can be made by multiplying the value (in *s*) of these quantities by $R_{link}/(8 \cdot 200)$. The average passive period, i.e. the average residence time in the passive state 'P', is then given by

$$T_P = \frac{1-\alpha}{\alpha} T_A \quad .$$

Both the active and the passive sojourn times are assumed to be geometrically distributed, with mean T_A and T_P respectively. The state transition diagram of such a MMBP source is depicted in Fig. III-20.

The entries of the 2×2 pgm $\mathbf{Q}(z)$ that completely characterises the packetised voice source described above, are then calculated as

$$\mathbf{Q}(z) = \begin{bmatrix} \left(1-\frac{1}{T_A}\right) G(z) & \frac{1}{T_A} \\ \frac{1}{T_P} G(z) & 1-\frac{1}{T_P} \end{bmatrix} \; ; \; G(z) = 1 + I_a^{-1}(z-1) \quad . \tag{III.59a}$$

Therefore, a phone that does not use silence suppression is completely described by three parameters : the (average) inter-packet time I_a, the average call duration T_A and the activity

grade α. The average number of packets per slot that one such a phone feeds to the node is equal to α/I_a; hence, the number of phones that is associated with a load ρ will equal

$$N=\rho I_a/\alpha \quad . \tag{III.59b}$$

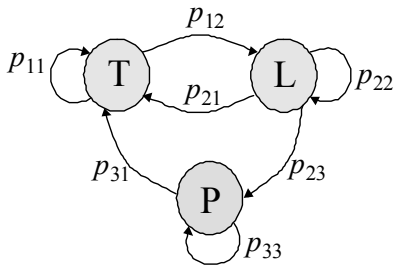


Figure III-21 : state transition diagram for a phone with silence suppression.

a phone with silence suppression

A phone that uses silence suppression can also be in an active 'A' or a passive 'P' state, but in the active state there are now two sub-states. An active phone with silence suppression can either be 'talking' or 'listening'. According to [186], the length of an average talking and listening period is about $1s$ and $1.5s$, respectively. In the former state, referred to as 'T', the phone generates packets as a Bernoulli source with the same rate $1/I_a$ as before, while in the latter state, referred to as 'L', as in the passive state 'P', no packets are generated at all.

Again, the sojourn times in the 'T', 'L', and 'P' states are geometrically distributed, with mean $T_T=1s$, $T_L=1.5s$ and $T_P=(1-\alpha)/\alpha\cdot 120s$ respectively. The state transition diagram of such a MMBP source is depicted in Fig. III-21. For this source modelling scheme, one can verify that the entries of the pgm $\mathbf{Q}(z)$ are now given by

$$\mathbf{Q}(z)=\begin{bmatrix} \left(1-\frac{1}{T_T}\right)G(z) & \frac{1}{T_T} & 0 \\ \left(\frac{1}{T_L}-\frac{1}{(1-\sigma)T_A}\right)G(z) & 1-\frac{1}{T_L} & \frac{1}{(1-\sigma)T_A} \\ \frac{1}{T_P}G(z) & 0 & 1-\frac{1}{T_P} \end{bmatrix} \; ; \; G(z)=1+I_a^{-1}(z-1) \quad , \tag{III.60a}$$

with the silence suppression factor σ defined as

$$\sigma=\frac{T_T}{T_T+T_L} \quad ,$$

and equal to 0.4 (=1/(1+1.5)) in this traffic scenario.

A phone that uses silence suppression is completely described by five parameters: the (average) inter-packet time I_a, the average call duration T_A, the activity grade α, the average

duration of a talk spurt T_T and of a silence period T_L. The load that a single phone with silence suppression imposes on the network node is now equal to $\sigma\alpha/I_a$; consequently, the number of phones associated with a (aggregate) load ρ is given by

$$N = \rho I_a / (\sigma\alpha) \quad . \tag{III.60b}$$

III.6.2.2 the burst factor

Based on these parameter settings, we can calculate the burst factor κ_b that corresponds to these packetised voice sources with and without silence suppression. First, for the case without silence suppression, the packet arrival process is basically an IBP, and if we carry out similar calculations as the ones leading to (III.31a), we can show that κ_b satisfies (see also [99])

$$\kappa_b \cong 1 + 2\frac{(1-\alpha)^2}{1-\rho/N}\left(\frac{T_A}{I_a}\right) \quad , \tag{III.61a}$$

where we have taken into consideration that, under normal circumstances, we have that $(1\text{-}\alpha)T_A >> 1$ since the average active period (of $120s$) is typically very long compared to the slot length (tens up to hundreds of μs), and retained the most significant term in the power series of $((1\text{-}\alpha)T_A)^{-1}$ that shows up in the exact expression for the burst factor. The second term in the right hand side accounts for the burstiness of the IBP, and the value of κ_b approaches 1 (i.e., the burst factor of an i.i.d. process) as $\alpha \to 1$; this corresponds to the intuitive notion that, for $\alpha \to 1$, the corresponding MMBP source almost always resides in the 'A' state and therefore resembles an i.i.d. Bernoulli process.

Following an analogous approach for the scenario with silence suppression, we find the following approximate formula for the burst factor

$$\kappa_b \cong 1 + 2\frac{(1-\alpha)^2\sigma}{1-\rho/N}\left(\frac{T_A}{I_a}\right) + 2\frac{(1-\sigma)^2}{1-\rho/N}\left(\frac{T_T}{I_a}\right) \quad . \tag{III.61b}$$

Observe that for $\sigma \to 1$ (which corresponds to a source that, when active, almost always resides in the 'T' state), this expression for κ_b becomes equal to (III.61a), in which case there is no silence suppression. Also note the similar roles that are played by σ and T_T in the third term in the right-hand side of this equation, compared to α and T_A in the right-hand side of (III.61a).

III.6.2.3 determining the tolerable load

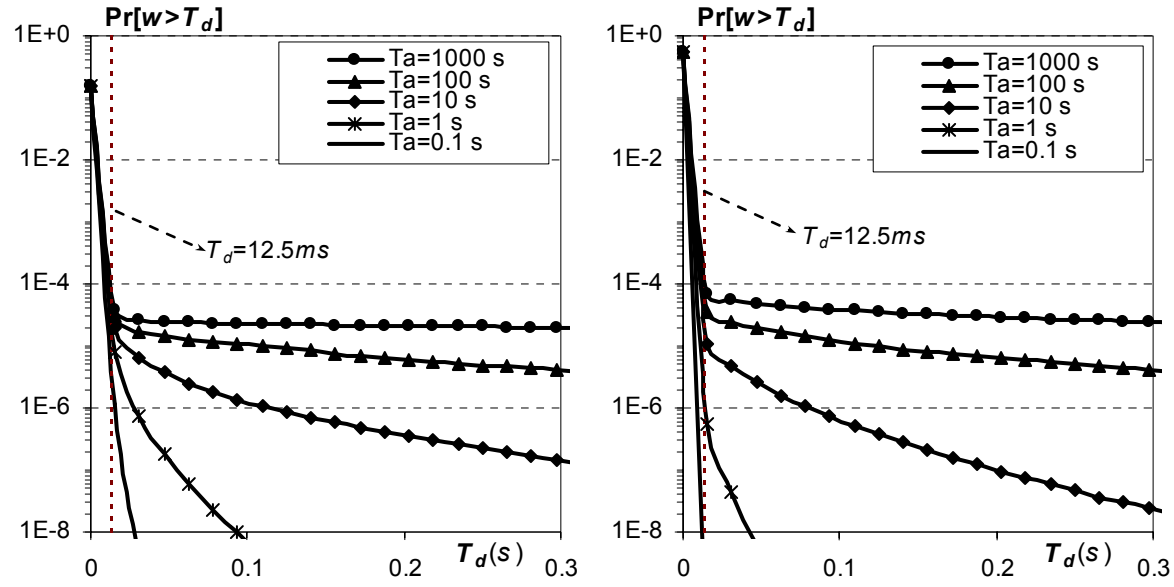
Since the VoIP sources, as described in the previous section, are described by a MMBP with either 2 or 3 states – depending on whether or not silence suppression is implemented – and

the packet transmission times equal 1 slot, the results of this chapter can be applied to assess the performance of the bottleneck access node. In particular, for a variety of traffic scenarios, we will determine the maximum value of N (or, equivalently, ρ) such that the bound T_d on the packet waiting time is exceeded by no more than a fraction of 10^{-X} packets, i.e.,

$$N_{\max} = \max_{N} \{\Pr[w > T_d]\} < 10^{-X} \quad ,$$

and the value of the tolerable load ρ_{max} then readily follows from the relations (III.59a) or (III.60b). In view of the previous remarks, we set T_d=12.5ms and X=4, unless mentioned otherwise.

From the proportionality between the burst factor κ_b and the mean active period T_a that follows from (III.61a) (and (III.61b) as well), one could be under the impression that a primordial role will be played by T_a in the calculation of ρ_{max}. This however, turns out to not to be the case : although there is a distinct (proportional) relation between the burst factor of the D-BMAP arrival process and the mean and variance of the packet waiting time, the impact of κ_b (and the parameters that determine κ_b) on the tail behaviour of the packet waiting pmf and/or ccdf is far from obvious. To illustrate this, we refer to the results that are plotted in Figs. III-22a,b for Pr[w>T_d], in case of phones without silence suppression, and values of R_{link}, R_{cod}, and the activity grade α as indicated. The values of the load ρ of 0.2 (a) and 0.6 (b) have been picked in such a way that Pr[w>12.5ms] is approximately equal to, but slightly less than, 10^{-4}, for T_a=120s. These results for the waiting time ccdf reveal that, due to this tail behaviour that is typical for the system under consideration, the moments of w will keep increasing as T_a becomes larger (and will indeed tend to infinity if T_a does so, in view of the 'horizontal' tail of the ccdfs in such a case), whereas this conclusion does not hold with respect to the waiting time quantiles. Indeed, as soon as T_a assumes a value that is sufficiently large (say tens of seconds), these plots indicate that a further increase of the load ρ will lead to a value of Pr[w>T_d] that will be larger than 10^{-4}, independent of the specific value of T_a (unless the bound T_d is chosen unacceptably large, a situation that we want to avoid). Hence, for the range of values of T_a and T_d that are of interest to us, the value of ρ_{max} will be virtually independent of T_a, in spite of the proportionality of the burst factor to T_a. These observations also make clear that, for the D-BMAP under consideration, it will not be possible to establish a simple (approximate) relation between the quantile of the packet waiting time and its mean and variance, as opposed to the i.i.d. case; see equation (II.50b) in Section II.4.4. In addition, this tail behaviour once more exemplifies the need for models that are highly accurate, since relatively small deviations in the system dimensioning may lead to considerable performance degradation.



(a) R_{link}=1.024 *Mb*/*s*, ρ=0.2 (b) R_{link}=10.24 *Mb*/*s*, ρ=0.6

Figure III-22 : tail behaviour of the packet waiting time ccd; R_{cod}=64 *kb*/*s*, α=0.1

In the numerical results that follow hereafter, we consider both the case of homogeneous and heterogeneous traffic. In the homogeneous case, all phones that compete for the available resources in one node use the same codec, and either all use silence suppression, or none do. In the heterogeneous case we consider a mix of traffic generated by phones that use one of two types of codecs, implying that we can identify two traffic classes. For completeness, let us also mention that we adopt the FOC ordering paradigm in case of heterogeneous traffic, and require that the stochastic bound of the above equation that is imposed on the packet waiting time must be satisfied for both classes.

For the homogeneous case, we also compare the tolerable load according to the MMBP/*D*/1 model with the one derived from the *M*/*D*/1 model, which neglects the correlation in the packet arrival pattern. Remark that once the tolerable load has been identified through the *M*/*D*/1 model, equations (III.59b) and (III.60b) can still be applied to calculate the number of phones that can be supported by the *M*/*D*/1 node for phones without and with silence suppression, respectively. We expect the tolerable load obtained with the *M*/*D*/1 model to give an upper bound for the one obtained with the MMBP/*D*/1 model, since an aggregate of such MMBP sources is burstier than a set of Bernoulli sources. This burstiness stems from the observation that a MMBP source generates packets at full rate during a time period that corresponds to a talk spurt, while for the *M*/*D*/1 model, packet arrivals are generated in a random manner (i.e., according to a Poisson process) at a rate that corresponds to the load ρ, and are therefore expected to be more equally spread in time.

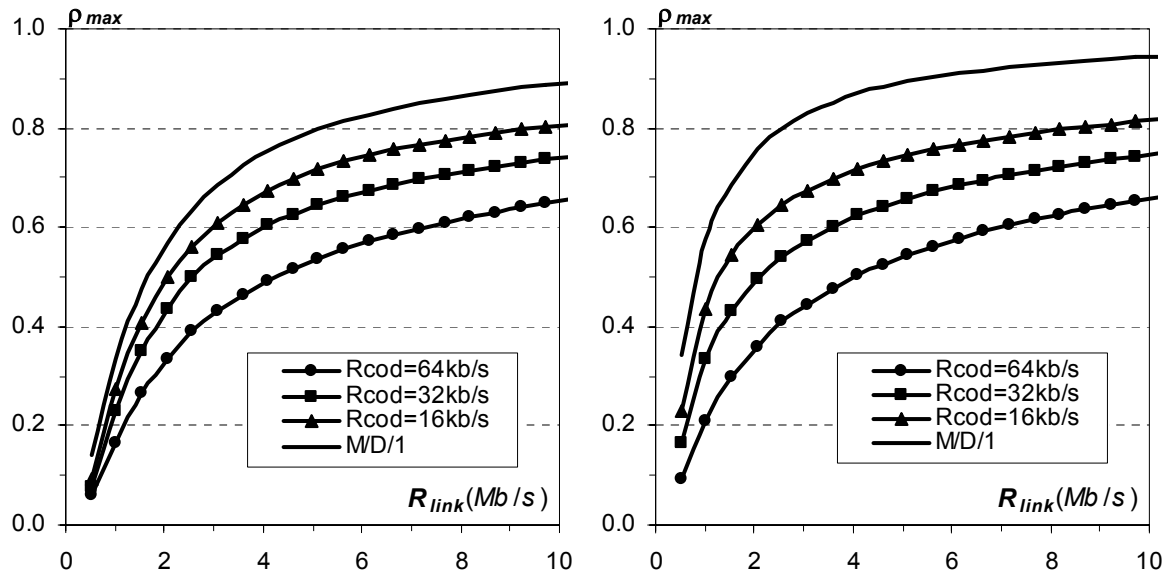


Figure III-23 : tolerable load, versus R_{link}, for phones without silence suppression; α=0.1

III.6.3 results

III.6.3.1 the influence of the codec bit rate R_{cod}

Figs. III-23a,b illustrate the effect of the codec bit rate R_{cod} on the tolerable load $\rho_{\max}$, versus R_{link}, for T_d=6.25ms (a) and T_d=12.5ms (b) respectively. Apparently, the tolerable load increases as R_{link} increases, which points to the fact that the network resources are used in an efficient manner only if the available bandwidth R_{link} is sufficiently high. As the value of R_{link} further increases, the tolerable load reaches a limiting value, and finding an efficient way to determine this value could be an interesting research topic as well, but has not been further pursued at this point. Note for instance that, as R_{link} increases further and further, the slot length, and hence the packet transmission times (expressed in seconds), converge to 0, while quantities such as the average length of an active or inactive period (expressed in slots) become infinitely long.

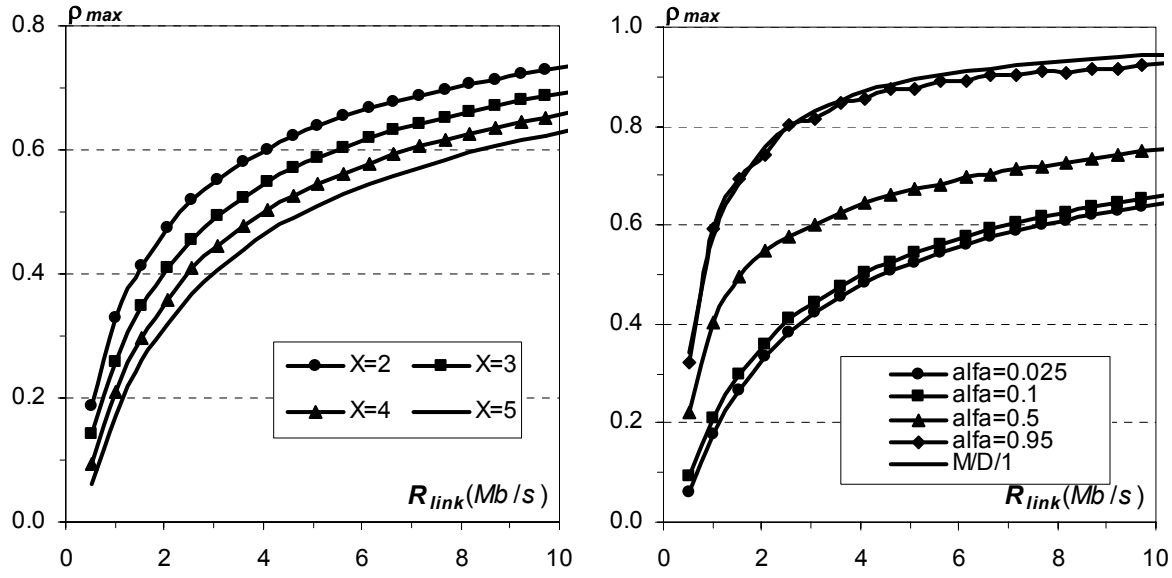
In addition, we also observe that the results that were depicted in Figs-III-23a,b indicate that the $M/D/1$ model overestimates the tolerable load, in some cases even by a factor 3, which highlights the need to use more sophisticated models than $M/D/1$ in the performance assessment of this kind of telecommunication network scenarios.

These figures reveal as well that the lower the bit rate of the codec, the larger the tolerable load on the node is. This is corroborated by expression (III.61a) for the burst factor κ_b, which shows that the smaller the codec rate (i.e. the larger I_a), the closer an aggregate of MMBP sources approximates an aggregate of Bernoulli sources, and hence, a Poisson process. We

would also like to point out that when considering fixed values of the load ρ, then since I_a is inversely proportional to R_{cod}, so is N, i.e., the lower the codec bit rate, the larger the number of phones that can be supported. Hence, on top of this gain that is inherent to using a low bit rate codec, an additional gain is achieved, because the node can be loaded up to a higher value. For example, for T_d=12.5ms and α=0.1 (see Fig. III-23a), the tolerable load for a link rate of 5.12Mb/s is 0.545 and 0.745 for the 64kb/s codec and the 16kb/s codec, respectively. The number of phones that can be supported for the codec with bit rate 16kb/s, is 1908, while it is only 349 for the codec with bit rate 64kb/s. This results in a total gain of 5.47 (=4·(0.745/0.545)) instead of the factor 4 that is anticipated, when the same delay quantile (of 12.5ms) has to be respected. We have to bear in mind as well, however, that the packet size for each codec bit rate was chosen the same (i.e. 200B). This means that the packetisation delay for a codec with a bit rate of 16kb/s is 4 times higher than the one for the codec with bit rate 64kb/s. If the packetisation delay would have been chosen the same for codecs with different bit rate, the filling factor φ for the 64kb/s codec would have been (a lot) larger than for the 16kb/s codec. From expression (III.59b) we can deduce that this filling factor φ too has a direct, and negative, impact on the number of phones that can be supported if the packet overhead increases. Since it is beyond the scope of this limited study to identify the optimal packetisation delay, it is not quantitatively investigated here what the impact is of an increase in packet waiting time budget T_d due to the fact that less packetisation delay is consumed keeping the mouth-to-ear delay budget equal to 150ms. Remark that also the time unit and serialization delay change as well, if the packet size changes.

A comparison of Figs. III-23a and III-23b illustrates that if the packet waiting time constraint is relaxed (i.e., T_d is increased from 6.25ms to 12.5ms), the tolerable load, and hence, the number of sources, increases, in particular for values of R_{link} that fall in the lower range. Also, and not unexpectedly, the tolerable load increases if the bound on the packet waiting time constraint is relaxed, i.e., if X decreases, as can be viewed in Fig. III-24a. However, the flip side of this trend is a decrease of conversation quality, due to the late arrival of an increasing fraction of packets.

Finally, for the sake of comparison, we would like to point out that, for R_{cod}=64kb/s, R_{link}=5.12Mb/s, α=0.1 and T_A=120s, then according to the (inverse) Engset formula, N=540 phones can be supported with a call blocking probability of 10^{-4} in the circuit-switched case, as opposed to N=349 that follows from the above D-BMAP. Hence, if the system dimensioning were to be merely based on the Engset queuing model, that is a bufferless model which does neglect the actual queuing phenomena, then the value for ρ_{max} that one thus obtains would be far too high. In view of the discussion that accompanied Figs. III-22a,b, we must conclude that this would lead to a considerable underestimation of the packet waiting times – or the packet loss due to buffer overflow if the buffer size were small compared to a



(a) impact of the bound 10^{-X}; α=0.1 (b) impact of the activity grade α; X=4

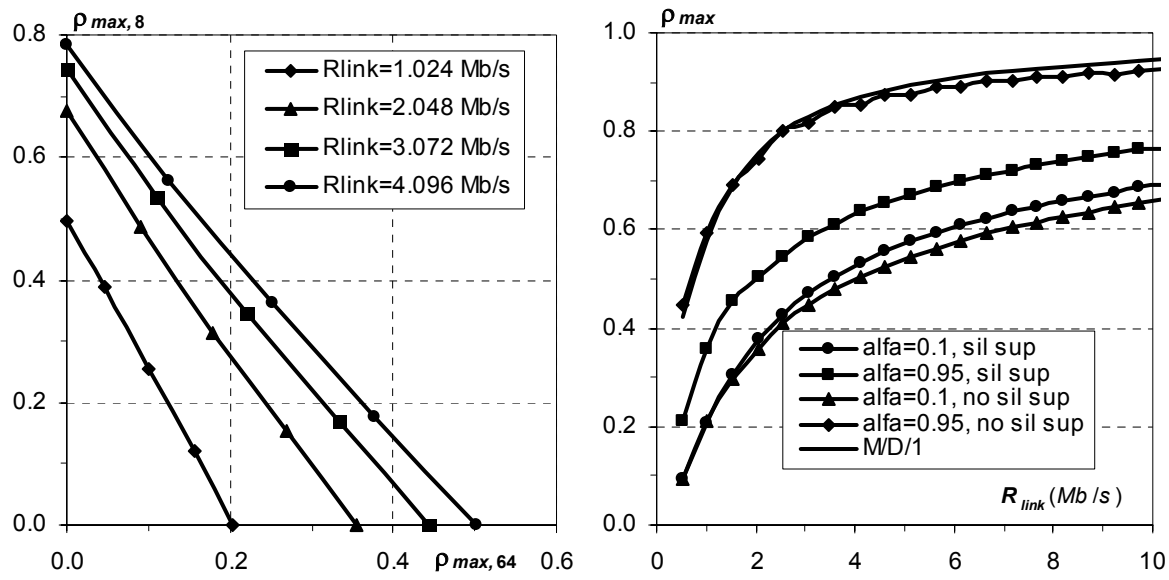
Figure III-24 : tolerable load, versus R_{link}, for phones without silence suppression; T_d=12.5 *ms*

typical waiting time that would be encountered in an infinite-size buffer – in the network node, thereby inducing a severe performance degradation in comparison to the preset targets. This also demonstrates that the constraint that we impose on the waiting time of the individual packets in the packet-switched case is more severe compared to the constraint on the call blocking probability in the circuit-switched case.

III.6.3.2 the influence of the activity grade α

The influence of the activity grade α for phones without silence suppression is shown in Fig. III-24b. It can be concluded that the higher the activity grade α, the better the *M*/*D*/1 model matches the MMBP/*D*/1 model. For α=0.95 (i.e., all phones are nearly always active) an MMBP source closely resembles a Bernoulli source, since during 95% of the slots, a packet arrival process is a Bernoulli process with rate $1/I_a$. Since an aggregation of Bernoulli sources, in turn, can be closely approximated by a Poisson process (provided that N is not too low), the *M*/*D*/1 and MMBP/*D*/1 models are (nearly) equivalent in this case.

There is not a lot of difference in the tolerable load when the activity grade α decreases from 0.1 to 0.025. Hence, supporting a number of phones that are active 10% of the time is more or less equivalent to supporting 4 times this number of phones that are active only 2.5% of the time. Therefore, if the activity grade is low enough, the number of phones that can be supported by the node is inversely proportional to the activity grade. However, if the activity grade takes values from 50% up to close to 100% (and in case no silence suppression is used), the number of phones that can be supported by the node is quite larger than suggested by this inverse proportionality rule.



(a) heterogeneous sources; α=0.1 (b) silence suppression; R_{cod}=64*kb/s*

Figure III-25 : tolerable load as a function of R_{link}; T_d=12.5*ms*, X=4

III.6.3.3 heterogeneous sources

In Fig. III-25a we consider the case of a heterogeneous traffic mix, where no silence suppression is applied. Again, we have set the activity grade α=0.1. In this figure we have plotted the tolerable load for a combination of two traffic classes that are multiplexed in a network node, for various values of the link rate R_{link}. Packets of the first traffic class are generated by 64*kb*/*s* codecs, while packets of the second traffic class are generated by 8*kb*/*s* codecs. We have taken the two extremes (64*kb*/*s* and 8*kb*/*s* codecs) of the codec bit rate considered in this case study to maximally illustrate the effect of mixing traffic with divergent characteristics. Since the packet size is the same (i.e. 200*B*) for both codecs, class 1 packets are generated every 20*ms* on average, while class 2 packets are generated every 160*ms* on average during an active period. The tolerable load (denoted by ρ_{64} and ρ_8 respectively) of both classes is again calculated by requiring that the fraction of packets of both classes having a queuing delay larger than 12.5*ms* may not exceed 10^{-4}.

From this figure we again deduce that little multiplexing gain is to be achieved when the link rate R_{link} is low. Only when R_{link} is at least a couple of *Mb*/*s*, the tolerable load of both classes reaches acceptable values. Secondly, it becomes once more clear that the lower bit rate codecs have a higher multiplexing gain (at the expense of a larger packetisation delay) as explained before. Most importantly, in view of the quasi-linear relation between the tolerable load combinations of both traffic classes, these curves could also have been easily and accurately derived from a linear interpolation of the results in the homogeneous case for 64*kb*/*s* and 8*kb*/*s* codecs respectively. This (close-to-linear) behaviour is (partly) caused by

the slot-based FCFS service discipline under consideration regardless of the traffic class that a packet belongs to, which implies that there is little or no discernable difference between the packet waiting time performance of the two classes, as explained in Section III.5.3.

III.6.3.4 the influence of silence suppression

In Fig. III-25b, we have summarised some results for the case where the silence suppression feature is switched on. First, we consider the case of the activity grade α=0.1, and compare the curves that correspond to the scenarios with and without silence suppression. We observe that the tolerable load is about the same irrespective of the fact whether silence suppression is used or not. Hence, the comparison of (III.59b) and (III.60b) reveals that there is a gain of $1/\sigma$ (i.e., 1/0.4=2.5 for this particular traffic scenario) that can be achieved in terms of the number of phones that can be supported, when silence suppression is switched on for all phones. The reason is that a value of α=0.1 already corresponds to a bursty source, because it injects packets into the network node at a full rate of R_{cod} for only 10% of the time, and remains passive for the rest of the time. The silence suppression mechanism then merely has a marginal – and even smoothing – effect on the buffer behaviour, as can also be deduced from comparison of expressions (III.61a,b) for the burst factor κ_b for both scenarios.

Next, we consider the case of phones that are nearly always active, i.e. α=0.95. Comparing the relevant curves on Fig. III-25b leads to the conclusion that switching on silence suppression does have a significant effect on the tolerable load in this case. When silence suppression is switched on, the network cannot be loaded up to the same value as in the case without silence suppression. For example for a link rate R_{link} of 5.12Mb/s the node can be loaded up to 0.89 when no silence suppression is used. When silence suppression is used the tolerable load drops to 0.63. Hence, the naively expected gain of 2.5 (=$1/\sigma$) is for some part undone by the increased burstiness of the sources with silence suppression, and drops to a value of approximately 2. Again, this qualitative behaviour can predicted by investigation of the dependence of expression (III.61b) for the burst factor κ_b on α and σ.

III.6.4 concluding remarks

We applied the results that were established in this chapter for the MMBP/D/1 queuing model, to determine the number of packet-based phones that can be supported by a multiplexing node.

We considered the case of homogeneous traffic generated by phones that either all use silence suppression or none of them do, as well as the case of heterogeneous traffic where the traffic is a mix generated by phones that use one of two possible bit rates. We investigated the impact of the activity grade, the codec bit rate and the link rate on the number of phones that can be supported by the packet-based node. We came to the following conclusions.

The $M/D/1$ model may considerably overestimate the number of sources that can be supported, even for large link rates where lots of phones can be supported. Apparently, an aggregate of this type of packetised voice sources is burstier than a Poisson process.

If the activity grade is considerably smaller than 100%, then the tolerable load is more or less independent of the activity grade, which leads to the rule that the number of phones that can be supported by the node is inversely proportional to the activity grade. However, if the activity grade approaches 100 % (in the case no silence suppression is used), the $M/D/1$ model is a better approximation, the tolerable load increases, and hence, the number of phones that can be supported by the node is larger than suggested by this inverse proportionality rule.

The use of a low bit rate codec leads to a considerable gain in the number of phones that can be supported by the network node. The gain is larger than just the codec gain, because an aggregate (of the same total bit rate and of the same packet size) becomes less bursty as the codec bit rate of the contributing sources decreases. However, we have to keep in mind that more packetisation delay is introduced for a low bit rate codec.

When mixing traffic with different profiles, the combinations of tolerable loads of the individual traffic classes practically follow a linear law. Hence, this observation provides an easy rule to calculate the tolerable load combinations in the case of heterogeneous traffic, if the tolerable loads in case of homogeneous traffic are known.

In addition, we observed that the bit rate reduction introduced by silence suppression largely outweighs the possible decrease in tolerable load that may be introduced by the additional burstiness of the aggregate traffic (e.g. if the activity grade is close to 1).

Finally, we would like to highlight the efficacy of the numerical procedures that were introduced throughout this chapter to calculated the performance indices of the buffer behaviour. Note for instance that the data point for ρ_{max} in Fig. III-25b for R_{link}=10.24Mb/s and α=0.1, corresponds to N=2217 packetised voice sources in case silence suppression is applied. Since this point has been calculated using a three-state D-BMAP for each source, expression (III.9b) reveals that this corresponds to a state space with 2460871 different states that can be visited by the Markov chain that determines the packet arrival process. First, it should be clear that the exact computation of the boundary probabilities, by solving a set of the same number of linear equations, is downright impossible. Also, the number of poles (and residues) that must be calculated in the multiple-pole tail approximation for the waiting time ccdf is of the same order of magnitude. Notwithstanding the huge size of this state space, we are still able to calculate the performance measures in a speedy and accurate manner.

III.7 Conclusions and unresolved issues

In this chapter, we have studied a multi-server queuing model with single-slot packet transmission times, where packet arrivals were generated by N ($\geq$1) D-BMAPs that were initially assumed to be identical. Considering an infinite-capacity buffer, then via a rather particular solution technique, we have investigated the queue and system content, and closed-form expressions for their steady-state pgfs were derived. This allowed us to establish (semi-)analytic closed-form formulae for the mean and variance of these quantities, which were expressed to the largest possible extent in terms of the traffic parameters. This has led to the insight that the so-called *burst factor* κ_b plays a central role in the performance assessment of this kind of queuing systems. Moreover, our solution approach has also led to the *multiple-pole* approximation technique for the tail behaviour of the queue and system content ccdf. In addition, we were able to establish a close relationship between the packet waiting time and delay pmf and pgf on the one hand, and the queue and system content pmf and pgf on the other hand, which is valid for *any kind* of packet arrival process.

Finally, considerable attention was devoted to the extension of these results to the case of a *heterogeneous* D-BMAP, where the results for the queue and system content pgfs bear a close resemblance to the homogeneous case. As far as the calculation of the packet waiting time and delay were concerned, we considered two *ordering schemes*, AO and FOC, and it became apparent that generating results for the moments and tail distribution in the former case can be hindered by considerable computational obstacles; in the FOC case however, results can be generated without great additional difficulty compared to the homogeneous packet arrival scenario.

We concluded this chapter with a case study for a scenario of packetised telephone traffic that feeds an output buffer in a network node, with the purpose of acquiring a better understanding of such a system, as well as illustrating the efficacy and usefulness of the results that were derived thus far.

In the course of this chapter, we have also run into a number of issues that are still unresolved or that may need some additional attention :

- due to the huge size of the state space that we occasionally encounter, we needed to establish close approximations for boundary probabilities, or, more precisely, the terms in our expressions for the mean, variance and tail distribution that contain the boundary probabilities. Especially in the multi-server case, a refinement of the results that have been obtained thus far would be useful;
- in the course of our analysis and the examples that accompany them, we encountered some intriguing properties concerning the buffer behaviour that may warrant further examination, such as the asymptotic behaviour of the buffer content and packet sojourn

time quantiles for increasing values of the burst factor κ_b; the (asymptotic) behaviour in our case study of the packet waiting time quantiles as the link rate becomes higher and higher;

- in case of an i.i.d. packet arrival process, we were able to establish close and useful relationships between the quantities of interest in the case of a buffer with finite storage capacity on the one hand, and an infinite-sized buffer on the other hand. We have not been able (yet) to derive similar relations for a buffer with a D-BMAP, and although some preliminary work has been done on this topic (e.g. [77]), it has not been fully resolved up to now.

Chapter IV

Queuing models with a D-BMAP and general packet transmission times

IV.1 Preface

In the previous chapters, we have focussed on a variety of multi-server output buffer models, which had a constant packet transmission time, equal to one slot, in common. In this chapter we extend these analyses to a queuing model with a generally distributed packet length and a correlated D-BMAP that generates packet arrivals. We will consider the case of a homogeneous, as well as a heterogeneous D-BMAP. This kind of model will be useful in assessing the performance of a packet-switched communication infrastructure, where messages carried by the network may have a variable transmission time, such as the current Internet, where the packet size ranges from a few tens of bytes up to $1500B$ (or even larger, for a very small percentage of packets; e.g. [156]). The service time of a packet equals its transmission time and is proportional to the size of a packet, which is assumed to be a generally distributed (i.i.d.) discrete random variable in this chapter, whose pmf (and pgf) could depend on the traffic class that the packet belongs to in case of a heterogeneous D-BMAP. We confine ourselves to the single-server case (i.e., c=1); it is well-known that discrete-time multi-server buffer models with generally distributed service times are notoriously hard to solve ([165], [260]), and even for the case of i.i.d. arrivals, a closed-form solution for quantities such as the pgf of the buffer content and packet sojourn time has yet to be found.

Let us as usual represent by ρ the load of the system (i.e., the fraction of time that the transmission unit is occupied). Due to the infinite buffer capacity, the system will reach a stochastic equilibrium only if the condition $\rho<1$ is fulfilled, and we will assume that such is the case in the remaining sections.

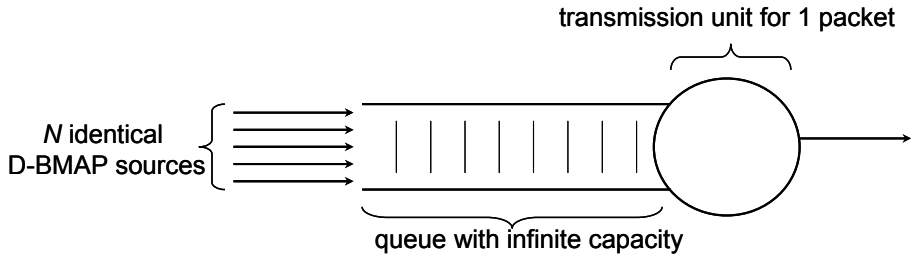


Figure IV-1 : infinite-capacity buffer with 1 output channel

IV.2 A homogeneous D-BMAP : analysis of the buffer content

We will study a discrete-time queuing model, as depicted in Fig. IV-1, that consists of one single server (also called transmission line, or output channel) and an infinite-capacity queue for storage of those packets that await transmission. Initially, we focus on the case where packet arrivals in the buffer are generated by N (identical, or homogeneous) sources that are each characterised by means of an L-state D-BMAP. Such a packet arrival model has been extensively described in Section III.2, and the notations, definitions and properties that were discussed therein will be unabatedly copied to the current section. Hence, the set $\overline{\boldsymbol{a}}_n = (a_{1,n}, \cdots, a_{L,n})$ of L drvs represents the numbers of sources that visit each of the L states, which constitutes a Markov chain that completely determines the packet arrival process, and is characterised by the pgm $\mathbf{Q}(z)$. At a later stage, this will be extended to the case where packet arrivals are generated by a heterogeneous D-BMAP.

In addition, packets that arrive in the buffer are of variable length. The transmission of a packet requires a positive integer number of slots, described by a general transmission time distribution, and can start (and end) at slot marks only, meaning that a packet transmission is synchronized with respect to the slot marks. It is assumed that the transmission times of consecutive packets that arrive in the buffer form a set of independent and identically distributed random variables, denoted by the drv t, with common pgf

$$T(z) \triangleq \boldsymbol{E}\left[z^t\right] = \sum_{l=1}^{\infty} z^l \Pr[t=l] \quad ,$$

where the average transmission time of a packet will be denoted by $\mu_t \equiv T'(1)$. Obviously, we assume that the transmission time of a packet requires at least one slot. For such a packet arrival pattern, it is obvious that the load ρ satisfies, in view of expression (III.11b) for μ_e

$$\rho = \mu_t \cdot \mu_e = \mu_t \cdot N \overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1) \overline{\boldsymbol{I}} \quad .$$

The analysis of this buffer model has been reported in [136] in case of an i.i.d. packet arrival process. An analysis of the discrete D-BMAP/G/1 – or a particular case, such as the

IBP, also called SBBP (switched batch Bernoulli process) – via the matrix-analytic or spectral decomposition technique has been reported in [206], [129], [176], and the references therein, while the finite-buffer case has been tackled in [130], [181], [251]. The analysis that is presented here extends the results of the previous chapter, and has been reported and applied in [82], [110], [114].

IV.2.1 the system equations

We first establish the system equations that control the evolution of the number of packets in the buffer during consecutive slots. As before, we let the drv s_n represent the system content at the beginning of slot n, which is the number of packets in the buffer at the beginning of slot n, including the one being transmitted (in case of a nonempty buffer), while e_n denotes the number of packet arrivals that are generated during slot n by the D-BMAP. In addition to s_n and the state of the Markov chain that controls the packet arrival process, we also need information about the amount of transmission time that the packet in service still requires at the beginning of slot n before being completely transmitted. We therefore define the *supplementary drv* τ_n as the *residual transmission time*, which is the number of slots, starting to count from the beginning of slot n, to complete the transmission of the packet that resides in the transmission unit in case of a nonempty buffer; when $s_n=0$, we automatically have that $\tau_n=0$. We can now establish the system equations that govern the buffer behaviour. We must distinguish between the three following cases :

1) $\underline{\tau_n=0 \ (\Leftrightarrow s_n=0)}$

When the buffer is empty at the beginning of a slot, the number of packets in the buffer at the beginning of the next slot equals the number of new arrivals, and we find that

$$s_{n+1} = s_n + e_n \quad . \tag{IV.1a}$$

If no packets have entered the buffer during slot n, the residual transmission time remains zero; otherwise it is equal to the transmission time of a new packet, which leads to

$$\tau_{n+1} = \begin{cases} 0 & \text{if } (e_n = 0) \\ t & \text{if } (e_n > 0) \end{cases} \quad . \tag{IV.1b}$$

2) $\underline{\tau_n=1}$

This implies that the packet in service is completely transmitted at the end of slot n and leaves the output buffer at this time instant

$$s_{n+1} = s_n + e_n - 1 \quad . \tag{IV.1c}$$

Similar to the previous case, if no packets have entered the buffer during slot n and the packet being transmitted was the only one in the buffer, then the buffer becomes empty and the residual transmission time equals zero at the beginning of the next slot, otherwise the transmission of a new packet commences, i.e.,

$$\tau_{n+1} = \begin{cases} 0 & \text{if } (e_n = 0) \text{ and } (s_n = 1) \\ t & \text{if } (e_n > 0) \text{ or } (s_n > 1) \end{cases} . \tag{IV.1d}$$

3) $\underline{\tau_n \geq 1}$

In this case, the packet being transmitted receives an extra slot of service without the transmission being completed, and we obtain

$$\begin{cases} s_{n+1} = s_n + e_n \\ \tau_{n+1} = \tau_n - 1 \end{cases} . \tag{IV.1e}$$

IV.2.2 derivation of a functional equation for the joint pgf of the state vector

Clearly, in view of the system equations (IV.1a-e), we need to keep track of the random variables s_n and τ_k in our state description, in order to capture the evolution of the system behaviour at (i.e., just after) consecutive slot boundaries. Consequently, due to the Markovian nature of the process that controls the number of packet arrivals during a slot, it becomes clear that the set of random variables $(s_n, \tau_n, \bar{\boldsymbol{a}}_{n-1})$ constitutes an (L+2)-dimensional state description of the system at the beginning of consecutive slots; this triplet will be referred to as the state vector in the remainder.

Let us now define the joint probability generating function of $(s_n, \tau_n, \bar{\boldsymbol{a}}_{n-1})$ as

$$P_{s,n+1}(z, y, \bar{\boldsymbol{x}}) \triangleq \mathbf{E}\left[z^{s_{n+1}} y^{\tau_{n+1}} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \right] = \mathbf{E}\left[z^{s_{n+1}} y^{\tau_{n+1}} \prod_{l=1}^{L} x_l^{\,a_{l,n}} \right] .$$

From system equations (IV.1a,b), and taking into account that τ_n=0 $\Leftrightarrow$ s_n=0, we then first of all derive that

$$\mathbf{E}\left[z^{s_{n+1}} y^{\tau_{n+1}} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \{s_n = 0, \tau_n = 0\} \right] = \mathbf{E}\left[z^{e_n} y^{t} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \{s_n = 0, \tau_n = 0\} \right] + \mathbf{E}\left[\left(y^0 - y^t \right) \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \{\tau_n = 0, s_n = 0, e_n = 0\} \right] .$$

In a similar way, equations (IV.1c,d) can be transformed into

$$\mathbf{E}\left[z^{s_{n+1}} y^{\tau_{n+1}} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \{\tau_n = 1\} \right] = \mathbf{E}\left[z^{s_n + e_n - 1} y^{t} \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \{\tau_n = 1\} \right] + \mathbf{E}\left[\left(y^0 - y^t \right) \bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_n} \{\tau_n = 1, s_n = 1, e_n = 0\} \right] .$$

Finally, (IV.1e) can be translated into a relation between z-transforms as well, yielding

$$\mathbf{E}\left[z^{s_{n+1}}y^{\tau_{n+1}}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\{\tau_n>1\}\right]=\mathbf{E}\left[z^{s_n+e_n}y^{\tau_n-1}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\right]$$
$$-y^{-1}\mathbf{E}\left[z^{e_n}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\{\tau_n=0,s_n=0\}\right]-\mathbf{E}\left[z^{s_n+e_n}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\{\tau_n=1\}\right].$$

Summation of the three foregoing equations then results in

$$yP_{s,n+1}(z,y,\overline{\boldsymbol{x}})=\mathbf{E}\left[z^{s_n+e_n}y^{\tau_n}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\right]+(yT(y)-1)\mathbf{E}\left[z^{e_n}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\{\tau_n=0,s_n=0\}\right]$$
$$+y(T(y)-z)\mathbf{E}\left[z^{s_n+e_n-1}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\{\tau_n=1\}\right]$$
$$+y\sum_{i=0}^{1}(1-T(y))\mathbf{E}\left[z^{s_n+e_n}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_n}\{\tau_n=i,s_n=i,e_n=0\}\right],$$

where we have incorporated the property that the drv t, which represents the length of a new packet whose transmission will commence at the beginning of slot n+1, depends neither on s_n, e_n or $\overline{\boldsymbol{a}}_n$, nor on the particular value of τ_n (i.e., 0 or 1). Similar to the corresponding derivations that were presented in the previous chapter, observe that, given that $\overline{\boldsymbol{a}}_{n-1}$ is known, then the value of $\overline{\boldsymbol{a}}_n$ (and the number of packet arrivals e_n) is not influenced by the actual value of the couple (s_n,τ_n). Therefore, by application of the identity (III.8), the previous relation can be converted into

$$yP_{s,n+1}(z,y,\overline{\boldsymbol{x}})=P_{s,n}(z,y,\mathbf{Q}(z)\overline{\boldsymbol{x}})+(yT(y)-1)\varphi_n(\mathbf{Q}(z)\overline{\boldsymbol{x}})+y(1-T(y))\varphi_n(\mathbf{Q}(0)\overline{\boldsymbol{x}})$$
$$+y(T(y)-z)\Upsilon_n(z,\mathbf{Q}(z)\overline{\boldsymbol{x}})+y(1-T(y))\Upsilon_n(0,\mathbf{Q}(0)\overline{\boldsymbol{x}}),$$

where

$$\begin{cases}\varphi_n(\overline{\boldsymbol{x}})\triangleq\mathbf{E}\left[\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_{n-1}}\{s_n=0,\tau_n=0\}\right]=\mathbf{E}\left[\prod_{l=1}^{L}x_l^{\,a_{l,n-1}}\{s_n=0,\tau_n=0\}\right]\\ \Upsilon_n(z,\overline{\boldsymbol{x}})\triangleq\mathbf{E}\left[z^{s_n-1}\overline{\boldsymbol{x}}^{\overline{\boldsymbol{a}}_{n-1}}\{\tau_n=1\}\right]=\mathbf{E}\left[z^{s_n-1}\prod_{l=1}^{L}x_l^{\,a_{l,n-1}}\{\tau_n=1\}\right]\end{cases}. \qquad \text{(IV.2a)}$$

First of all, due to the foregoing definitions, we have $P_n(0,y,\overline{\boldsymbol{x}})\equiv P_n(0,0,\overline{\boldsymbol{x}})=\varphi_n(\overline{\boldsymbol{x}})$. Consequently, if we insert z=0 into the above expression for $P_{s,n+1}(z,y,\overline{\boldsymbol{x}})$, we derive that

$$\varphi_{n+1}(\overline{\boldsymbol{x}})=\varphi_n(\mathbf{Q}(0)\overline{\boldsymbol{x}})+\Upsilon_n(0,\mathbf{Q}(0)\overline{\boldsymbol{x}}).$$

Let us now assume that the equilibrium condition is satisfied, implying that all the functions that occur in the two previous equations evolve to a steady-state limit, which will be reflected in the remainder by suppressing the subscript n that represented the time-dependence of the tagged quantities in the previous derivations. Invoking the previous relation, we thus obtain

$$yP_s(z,y,\bar{\boldsymbol{x}}) = P_s(z,y,\mathbf{Q}(z)\bar{\boldsymbol{x}}) + y(1-T(y))\varphi(\bar{\boldsymbol{x}}) + (yT(y)-1)\varphi(\mathbf{Q}(z)\bar{\boldsymbol{x}}) + y(T(y)-z)\Upsilon(z,\mathbf{Q}(z)\bar{\boldsymbol{x}}) \tag{IV.2b}$$

Equation (IV.2b), combined with definition (IV.2a), defines a functional equation for $P_s(z,y,\bar{\boldsymbol{x}})$, the steady-state joint pgf of the random variables $(s,\tau,\bar{\boldsymbol{a}})$, i.e., the system content and residual transmission time at the beginning of an arbitrary slot, and the state of the Markovian packet arrival process during the preceding slot. From this result, by an analogous calculation as in Section III.3.2, we can derive a closed-from expression for $P_s(z,y,\bar{\boldsymbol{x}})$ that enables us to establish expressions for the relevant performance measures concerning the buffer behaviour, such as the mean, variance and tail of the ccdf of the system and queue content.

IV.2.3 solution of the functional equation

If we substitute consecutive arguments $\bar{\boldsymbol{x}} = \mathbf{Q}(z)^h \bar{\boldsymbol{x}}, 0 \le h \le H-1$, into expression (IV.2b), multiply both hand sides by y^{-h-1}, and take the sum of the equations that result thereof, we find

$$P_s(z,y,\bar{\boldsymbol{x}}) = y^{-H} P_s\left(z,y,\mathbf{Q}(z)^H \bar{\boldsymbol{x}}\right) + (1-T(y)) \sum_{h=1}^{H} y^{-h+1} \varphi\left(\mathbf{Q}(z)^{h-1} \bar{\boldsymbol{x}}\right) + (yT(y)-1) \sum_{h=1}^{H} y^{-h} \varphi\left(\mathbf{Q}(z)^h \bar{\boldsymbol{x}}\right) + y(T(y)-z) \sum_{h=1}^{H} y^{-h} \Upsilon\left(z,\mathbf{Q}(z)^h \bar{\boldsymbol{x}}\right).$$

We once again decompose the powers of the pgm $\mathbf{Q}(z)$ in this expression in terms of its eigenvalues and –vectors as $\mathbf{Q}(z)^h = \mathbf{W}(z)\mathbf{\Lambda}(z)^h \mathbf{U}(z)$, and define the functions $F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z)$, $\bar{\boldsymbol{l}},\bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}}$, as in (III.17a). If we now let H approach infinity, and consider values of y and z for which $|\lambda_i(z)| < |y| \le 1$, $1 \le i \le L$, then the first term in the right hand side of the above expression, which can be written as

$$\lim_{H\to\infty} y^{-H} P_s\left(z,y,\mathbf{Q}(z)^H \bar{\boldsymbol{x}}\right) = \sum_{j=0}^{\infty} z^j \sum_{k=0}^{\infty} y^k \sum_{\bar{\boldsymbol{l}},\bar{\boldsymbol{m}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z) \Pr\left[s=j, h=k, \bar{\boldsymbol{a}}=\bar{\boldsymbol{l}}\right] \cdot \lim_{H\to\infty} y^{-H} \prod_{i=1}^{L} \left(\lambda_i(z)^H \bar{\boldsymbol{w}}_i(z) \bar{\boldsymbol{x}}\right)^{m_i},$$

will converge to zero.

For the same reasons as explained in Section III.3.2 we can show that such values of y and z exist; consider for instance $|y|=1$ and $z \in \boldsymbol{C}_1 = \{z : |z|=1 \wedge z \ne 1\}$. Defining the steady-state boundary probabilities $p(\bar{\boldsymbol{l}})$ and the unknown functions $\Upsilon_{\bar{\boldsymbol{l}}}(z)$ as

$$p\left(\bar{\boldsymbol{l}}\right) \triangleq \frac{1}{1-\rho} \lim_{n\to\infty} \Pr\left[s_n = 0, \tau_n = 0, \bar{\boldsymbol{a}}_{n-1} = \bar{\boldsymbol{l}}\right]$$
$$\Upsilon_{\bar{\boldsymbol{l}}}(z) \triangleq \frac{1}{1-\rho} \lim_{n\to\infty} \mathbf{E}\left[z^{s_n-1}\left\{\tau_n = 1, \bar{\boldsymbol{a}}_{n-1} = \bar{\boldsymbol{l}}\right\}\right] \quad , \qquad \text{(IV.3a)}$$

then by following a similar reasoning, as above, we can deduce that the three remaining terms in the right-hand side of the latter equation for $P_s(z,y,\bar{\boldsymbol{x}})$ for $H \to \infty$ converge to

$$\sum_{h=a}^{\infty} y^{-h} \varphi\left(\mathbf{Q}(z)^h \boldsymbol{x}\right) = (1-\rho) \sum_{\bar{\boldsymbol{m}}} \frac{y^{1-a}\left(\boldsymbol{\Lambda}(z)^a \mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}}{y - E_{\bar{\boldsymbol{m}}}(z)} \Psi_{\bar{\boldsymbol{m}}}(z) \quad ; \quad a = 0,1$$

$$\sum_{h=1}^{\infty} y^{-h} \Upsilon(z, \mathbf{Q}(z)^h \bar{\boldsymbol{x}}) = (1-\rho) \sum_{\bar{\boldsymbol{m}}} \frac{\left(\boldsymbol{\Lambda}(z)\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}}{y - E_{\bar{\boldsymbol{m}}}(z)} \Gamma_{\bar{\boldsymbol{m}}}(z) \quad ,$$

where

$$\Gamma_{\bar{\boldsymbol{m}}}(z) \triangleq \sum_{\bar{\boldsymbol{l}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z) \Upsilon_{\bar{\boldsymbol{l}}}(z) \; ; \; \Psi_{\bar{\boldsymbol{m}}}(z) \triangleq \sum_{\bar{\boldsymbol{l}}} F_{\bar{\boldsymbol{l}}\bar{\boldsymbol{m}}}(z) p\left(\bar{\boldsymbol{l}}\right) \; ; \; E_{\bar{\boldsymbol{m}}}(z) \triangleq \left(\boldsymbol{\Lambda}(z)\bar{\boldsymbol{l}}\right)^{\bar{\boldsymbol{m}}} \quad , \qquad \text{(IV.3b)}$$

in accordance to the definitions of (III.18a). Summarising, in view of these definitions and results, and taking the limit $H \to \infty$ in the previous equation for $P_s(z,y,\bar{\boldsymbol{x}})$, we thus obtain

$$P_s(z,y,\bar{\boldsymbol{x}}) = (1-\rho) \sum_{\bar{\boldsymbol{m}}} \frac{\left(\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}}{y - E_{\bar{\boldsymbol{m}}}(z)}$$
$$\left\{\left[y\left(T(y) - z\right)\Gamma_{\bar{\boldsymbol{m}}}(z) + \left(yT(y) - 1\right)\Psi_{\bar{\boldsymbol{m}}}(z)\right] E_{\bar{\boldsymbol{m}}}(z) + y\left(1 - T(y)\right)\Psi_{\bar{\boldsymbol{m}}}(z)\right\} \quad .$$

This expression for the steady-state joint pgf $P_s(z,y,\bar{\boldsymbol{x}})$ still contains the unknown constants $p\left(\bar{\boldsymbol{l}}\right)$, as well as the unknown functions $\Upsilon_{\bar{\boldsymbol{l}}}(z)$, or, equivalently, the functions $\Gamma_{\bar{\boldsymbol{m}}}(z)$ and $\Psi_{\bar{\boldsymbol{m}}}(z)$. Let us therefore, for each value of $\bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}}$, consider values of y for which $y = E_{\bar{\boldsymbol{m}}}(z)$, which will be represented by $y_{\bar{\boldsymbol{m}}}(z)$. Then, in view of $|\lambda_i(z)| \leq 1$ for $|z| \leq 1$, the inequality $|E_{\bar{\boldsymbol{m}}}(z)| \leq 1$ also holds for values of z that fall within the complex unit disk, and hence, $|y_{\bar{\boldsymbol{m}}}(z)| \leq 1$ for such values of z. Consequently, the joint pgf $P_s(z,y,\bar{\boldsymbol{x}})$ is a bounded function for these values of z and y. This means that, if we multiply both hand sides of the above relation by $y - E_{\bar{\boldsymbol{m}}}(z)$ and consider $y = y_{\bar{\boldsymbol{m}}}(z)$, then the left-hand side becomes equal

to zero, implying that the same must hold for the right-hand side. For each value of $\bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}}$, this leads to the following equation :

$$E_{\bar{\boldsymbol{m}}}(z)\left(T\left(E_{\bar{\boldsymbol{m}}}(z)\right)-z\right)\Gamma_{\bar{\boldsymbol{m}}}(z)+T\left(E_{\bar{\boldsymbol{m}}}(z)\right)\left(E_{\bar{\boldsymbol{m}}}(z)-1\right)\Psi_{\bar{\boldsymbol{m}}}(z)=0 \quad , \quad \bar{\boldsymbol{m}} \in \Omega_{\bar{\boldsymbol{a}}} \quad ,$$

which, in effect, defines $\Gamma_{\bar{\boldsymbol{m}}}(z)$ in terms of $\Psi_{\bar{\boldsymbol{m}}}(z)$ for each value of $\bar{\boldsymbol{m}}$. Therefore, if we insert this result into the above expression for $P_S(z,y,\bar{\boldsymbol{x}})$, we finally obtain

$$\boxed{P_S(z,y,\bar{\boldsymbol{x}})=(1-\rho)\sum_{\bar{\boldsymbol{m}}}\left(\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}.\left\{1+zy\frac{\left(E_{\bar{\boldsymbol{m}}}(z)-1\right)\left(T(y)-T\left(E_{\bar{\boldsymbol{m}}}(z)\right)\right)}{\left(y-E_{\bar{\boldsymbol{m}}}(z)\right)\left(z-T\left(E_{\bar{\boldsymbol{m}}}(z)\right)\right)}\right\}\Psi_{\bar{\boldsymbol{m}}}(z)} \quad . \text{ (IV.4a)}$$

This final result for $P_S(z,y,\bar{\boldsymbol{x}})$ expresses the joint pgf of the state of the arrival process during an arbitrary slot, and the buffer content and residual transmission time at the beginning of the following slot, in terms of the unknown functions $\Psi_{\bar{\boldsymbol{m}}}(z)$, and hence, the unknown boundary probabilities $p(\bar{\boldsymbol{l}})$. In the next section, it will become clear how these can be computed. An expression for $P_q(z,y,\bar{\boldsymbol{x}})$, the joint pgf of the queue content and residual service time at the beginning of an arbitrary slot, and the state of the D-BMAP during the foregoing slot, ensues from the relation $q=(s-1)^+$, yielding the equality

$$P_q(z,y,\bar{\boldsymbol{x}}) \triangleq \lim_{n\to\infty}\mathbf{E}\left[z^{q_n}y^{\tau_n}\bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_{n-1}}\right]=z^{-1}\left\{P_S(z,y,\bar{\boldsymbol{x}})+(1-\rho)(z-1)\sum_{\bar{\boldsymbol{m}}}\left(\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}\Psi_{\bar{\boldsymbol{m}}}(z)\right\} \quad ,$$

which produces the following result :

$$\boxed{P_q(z,y,\bar{\boldsymbol{x}})=(1-\rho)\sum_{\bar{\boldsymbol{m}}}\left(\mathbf{W}(z)\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}.\left\{1+y\frac{\left(E_{\bar{\boldsymbol{m}}}(z)-1\right)\left(T(y)-T\left(E_{\bar{\boldsymbol{m}}}(z)\right)\right)}{\left(y-E_{\bar{\boldsymbol{m}}}(z)\right)\left(z-T\left(E_{\bar{\boldsymbol{m}}}(z)\right)\right)}\right\}\Psi_{\bar{\boldsymbol{m}}}(z)} \quad . \text{ (IV.4b)}$$

It is not difficult to check, by setting $y=1$ and $T(z)=z$, that these formulae indeed reduce to the corresponding respective pgfs $P_S(z,\bar{\boldsymbol{x}})$ and $P_q(z,\bar{\boldsymbol{x}})$ for the case of single-slot service times, i.e., (III.18a,b) with $c=1$.

Also, starting from these results, we will be able to establish closed-form expressions for the steady-state pgfs of the queue and system content, and their corresponding moments and asymptotic tail behaviour.

To conclude, let us take a brief look at the steady-state pgf of the residual transmission time of an arbitrarily chosen packet. Once more making use of $\Psi_{\bar{\boldsymbol{m}}}(1)=0$ unless $\bar{\boldsymbol{m}}=\bar{\boldsymbol{N}}$

(and the normalisation condition $\Psi_{\bar{N}}(\bar{1})=1$; see also the remarks in the next section concerning this topic), we find that

$$T_r(y) \triangleq \mathbf{E}\left[y^{\tau}\right] = 1-\rho+\rho T_{r,0}(y) \; ; \; T_{r,0}(y) \triangleq y\frac{T(y)-1}{\mu_t(y-1)} \quad , \tag{IV.5a}$$

This equation, among others, reveals that the joint pgfs $P_s(z,y,\bar{\boldsymbol{x}})$ and $P_q(z,y,\bar{\boldsymbol{x}})$ are indeed normalised if we set $y=z=1$ and $\bar{\boldsymbol{x}}=\bar{\boldsymbol{I}}$. We would like to point out as well that the pgf $T_{r,0}(y)$, which represents the pgf of the drv τ_0 – denoting the residual transmission time in case the buffer is nonempty – can also be derived through a similar reasoning as was done in Section II.4.1 for the drv f, since the residual transmission time of an arbitrary packet is equal to the number of slots between a randomly chosen transmission slot of the packet, and the end of its transmission epoch. Exploiting this formal similarity between the drvs f and τ_0, with $T_{r,0}(z)=zF(z)$ we thus find for the moments of τ_0 in view of (II.38) :

$$\begin{aligned} \mu_{\tau_0} &= \frac{\sigma_t^2}{2\mu_t}+\frac{\mu_t+1}{2} \\ \sigma_{\tau_0}^2 &= \frac{\mu_{3,t}}{3\mu_t}-\left(\frac{\sigma_t^2}{2\mu_t}\right)^2+\frac{\sigma_t^2}{2}+\frac{\mu_t^2-1}{12} \end{aligned} \quad , \tag{IV.5b}$$

where $\mu_{3,t}$ represents the third central moment of the packet transmission time, and σ_t^2 its variance.

IV.2.4 calculation of the boundary probabilities

We can calculate the probabilities $p(\bar{\boldsymbol{l}})$, defined in (IV.3a), by expressing that the joint pgf $P_S(z,y,\bar{\boldsymbol{x}})$ is an analytic function when the complex variables y and z both lie inside the complex unit circle, i.e., $|z|\le 1$ and $|y|\le 1$. We have already expressed in the previous section that for $|y|\le 1$ and $y=E_{\bar{\boldsymbol{m}}}(z)$, $P(z,y,\boldsymbol{x})$ is finite. In view of expressions (IV.4a,b), the only remaining potential singularities inside the complex unit disk are those values of z (and y) for which $z=T(E_{\bar{\boldsymbol{m}}}(z))$. Indeed, although it is not possible to prove by means of Rouché's theorem, since the eigenvalues $\lambda_l(z)$, $1\le l\le L$, are not necessarily analytic inside the complex unit disk, it turns out that this equation has exactly one solution inside this region. Let us for each value of $\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}$, denote this solution by $z_{\bar{\boldsymbol{m}}}$, i.e.,

$$z_{\bar{\boldsymbol{m}}}=T(E_{\bar{\boldsymbol{m}}}(z_{\bar{\boldsymbol{m}}})) \; , \; |z_{\bar{\boldsymbol{m}}}|\le 1 \; , \; \bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}} \quad , \tag{IV.6a}$$

with $z_{\bar{N}} \equiv 1$. Then, by expressing that $z_{\bar{m}}$ must also be a zero of the corresponding numerator in expression (IV.4a) for $P_s(z, y, \bar{x})$, we obtain

$$\Psi_{\bar{m}}(z_{\bar{m}}) = 0 \quad , \ \bar{m} \in \Omega_{\bar{a}} \setminus \bar{N} \quad ; \quad \Psi_{\bar{N}}(1) = 1 \quad , \tag{IV.6b}$$

where the trivial equation that is associated with $z_{\bar{N}} = 1$ has been replaced by the normalisation condition, which yields $\Psi_{\bar{N}}(1) = 1$ in view of definition (IV.3a) and the property Pr[s=0]=1-ρ. This result can be deduced in a similar way as in Section II.3.1 (e.g. equation (II.8)) by regarding a packet of length l as a concatenation of l 'minipackets' of length 1 (i.e., with single-slot transmission times). These identities constitute a set of linear equations for the same number of unknowns – as explained in Section III.3.3 – that can be solved numerically if the size of this set remains within reasonable limits.

In case we want to altogether avoid the numerical computation of the boundary probabilities when deriving numerical results for the performance measures, we can adopt the same approach as in Section III.3.3, which once more yields approximation (III.21a) for the functions $\Psi_{\bar{m}}(z)$ that contain these quantities. In concurrence with the procedure that was outlined in that section, the parameter M that appears in these approximate formulae can now be determined as being the smallest integer for which $C_a^{-1} \leq 1-\rho$.

IV.3 The queue and system content

IV.3.1 derivation of the probability generating function

If we insert y=1 and $\bar{x} = \bar{1}$ into expressions (IV.4a,b), we obtain an expression for the steady-state pgfs $S(z)$ and $Q(z)$ of the system and queue content respectively :

$$S(z) = (1-\rho) \sum_{\bar{m}} \frac{(z-1) T(E_{\bar{m}}(z))}{z - T(E_{\bar{m}}(z))} \Psi_{\bar{m}}(z) \tag{IV.7a}$$

$$Q(z) = (1-\rho) \sum_{\bar{m}} \frac{(z-1)}{z - T(E_{\bar{m}}(z))} \Psi_{\bar{m}}(z) \quad . \tag{IV.7b}$$

From these expressions, one can easily derive the performance measures of interest related to the queue and system content, as shown in the subsequent sections.

IV.3.2 mean and variance

The mean and variance of the queue content q can be calculated by letting the $\mathcal{M}[\cdot]$ and $\mathcal{V}[\cdot]$ operators act upon the previous expression for $Q(z)$. First, consider the generic drvs f, g and h with pgfs $F(z)$, $G(z)$ and $H(z)$ respectively, and suppose that $H(z)=F(G(z))$. One can then show that

$$\begin{aligned}
\mu_h &= \mu_f \mu_g \\
\sigma_h^2 &= \mu_f \sigma_g^2 + \mu_g^2 \sigma_f^2 \\
\mu_{3,h} &= \mu_f \mu_{3,g} + 3\mu_g \sigma_f^2 \sigma_g^2 + \mu_g^3 \mu_{3,f}
\end{aligned} \quad .$$

Consequently, in view of the results that were generated in Section III.3.5.1, and the similar role that is now played by $T(E_{\bar{m}}(z))$, compared to $E_{\bar{m}}(z)$ in the previous chapter, it is rather straightforward to show that

$$\mu_q = \frac{\mu_t \sigma_e^2 \kappa_b + \mu_e^2 \sigma_t^2}{2(1-\rho)} - \frac{\rho}{2} + \mathcal{M}\left[\Psi_{\bar{N}}\right] \quad , \tag{IV.8a}$$

$$\mathcal{M}\left[\Psi_{\bar{N}}\right] = \sum_{\bar{l}} \left(\sum_{i=1}^{L} l_i u'_{i1}(1) \right) p(\bar{l})$$

and

$$\begin{aligned}
\sigma_q^2 = {} & \frac{\mu_t \mu_{3,e} \kappa_s + 3\mu_e \sigma_t^2 \sigma_e^2 \kappa_b + \mu_e^3 \mu_{3,t}}{3(1-\rho)} + \left(\frac{\mu_t \sigma_e^2 \kappa_b + \mu_e^2 \sigma_t^2}{2(1-\rho)} \right)^2 - \frac{\mu_t \sigma_e^2 \kappa_b + \mu_e^2 \sigma_t^2}{2} + \frac{1-(1-\rho)^2}{12} \\
& + \mathcal{V}\left[\Psi_{\bar{N}}\right] + 2 \sum_{\bar{m} \in \Omega_{\bar{a}} \setminus \bar{N}} \frac{(1-\rho)}{1-T(E_{\bar{m}}(1))} \mathcal{M}\left[\Psi_{\bar{m}}\right]
\end{aligned} \tag{IV.8b}$$

$$\mathcal{V}\left[\Psi_{\bar{N}}\right] = \sum_{\bar{l}} \left\{ \sum_{i=1}^{L} l_i \left(u''_{i1}(1) + u'_{i1}(1) - u'_{i1}(1)^2 \right) + \left(\sum_{i=1}^{L} l_i u'_{i1}(1) \right)^2 \right\} p(\bar{l}) - \mathcal{M}\left[\Psi_{\bar{N}}\right]^2 \quad ,$$

and where the last term in the right-hand side of (IV.8b) can be converted into

$$\sum_{\bar{m} \in \Omega_{\bar{a}} \setminus \bar{N}} \frac{(1-\rho)}{1-T(E_{\bar{m}}(1))} \mathcal{M}\left[\Psi_{\bar{m}}\right] = (1-\rho) \sum_{\bar{l} \in \Omega_{\bar{a}}} \left(\sum_{i=1}^{L} l_i \sum_{k=2}^{L} \frac{u'_{ik}(1)}{1-T(\lambda_k(1))} \right) p(\bar{l}) \quad .$$

In addition, the mean and variance of the system content can be expressed in terms of the mean and variance of the queue content in a similar way as in (III.29), leading to

$$\begin{aligned}
\mu_s &= \mu_q + \rho \\
\sigma_s^2 &= \sigma_q^2 + \mu_t \sigma_e^2 \kappa_b + \mu_e^2 \sigma_t^2 + 2(1-\rho) \mathcal{M}\left[\Psi_{\bar{m}}\right]
\end{aligned} \quad . \tag{IV.8c}$$

The heavy-load approximation for the mean and variance of the queue and system content now becomes

$$\mu_{q,\rho\to 1} = \mu_{s,\rho\to 1} = \frac{\mu_t \kappa_b \sigma_e^2 + \mu_e^2 \sigma_t^2}{2(1-\rho)}$$
$$\sigma^2_{q,\rho\to 1} = \sigma^2_{s,\rho\to 1} = \left(\frac{\mu_t \kappa_b \sigma_e^2 + \mu_e^2 \sigma_t^2}{2(1-\rho)} \right)^2 \quad , \tag{IV.8d}$$

and from these formulae, it becomes clear that the variance of the packet transmission times plays a similar role in the calculation of these quantities as κ_b and σ_e^2. Note however that with $\mu_t \geq 1$ we necessarily have that $\mu_e < 1$ due to the equilibrium condition $\mu_t \cdot \mu_e < 1$, and the larger the value of μ_t, the smaller the value of μ_e will be. Consequently, the numerator of the above expressions is likely to be dominated by the term $\mu_t \kappa_b \sigma_e^2$ – unless σ_t^2 assumes extremely high values – which shows that the system performance is still largely dominated by the burstiness of the packet arrival process under these circumstances. Also, these expressions illustrate that it is still so that the coefficients of variation of both the queue and system content evolve to 1 for $\rho \to 1$.

Finally, if we wish to shun the explicit numerical computation of the boundary probabilities, we can invoke the approximation (III.21a) for $\Psi_{\bar{\boldsymbol{m}}}(z)$, which also in this case leads to expressions such as (III.30a,b), where we need to replace $E_{\bar{\boldsymbol{m}}}(1)$ and $\lambda_k(1)$ by $T(E_{\bar{\boldsymbol{m}}}(1))$ and $T(\lambda_k(1))$ respectively in both hand sides of (III.30b).

IV.3.3 tail distribution

The ccdf of the queue and system content can be approximated by adopting a similar multiple-pole tail approximation as in Section III.3.6.1, where the poles that determine this tail behaviour are now the solutions of the equations

$$z_{o,\bar{\boldsymbol{m}}} = T\left(E_{\bar{\boldsymbol{m}}}\left(z_{o,\bar{\boldsymbol{m}}}\right)\right) \equiv T\left(\prod_{j=1}^{L} \lambda_j \left(z_{o,\bar{\boldsymbol{m}}}\right)^{m_j} \right) \; ; \; z_{o,\bar{\boldsymbol{m}}} \in]1, \mathcal{R}[\; , \; \bar{\boldsymbol{m}} \in \Omega_{o,\bar{\boldsymbol{a}}} \quad . \tag{IV.9a}$$

If we let $\mathcal{R}_t$ and $\mathcal{R}_q$ represent the radii of convergence of $T(z)$ and the pgm $\mathbf{Q}(z)$ respectively, then the radius of convergence $\mathcal{R}$ is presently equal to the minimum of $\mathcal{R}_q$, and the value z^* defined by $E_{\bar{\boldsymbol{m}}}(z^*) = \mathcal{R}_t$ (if such a solution exists for $z^* < \mathcal{R}_q$). These equations will have a solution for those values of $\bar{\boldsymbol{m}}$ for which the inequality

$$\lim_{\substack{z \to \mathcal{R} \\ <}} T(E_{\bar{\boldsymbol{m}}}(z))/z > 1 \quad ,$$

holds, and we let $\Omega_{o,\bar{a}}$ represent this set of values. Thus, with the coefficients $\zeta_{\bar{m}}$ of $Q(z)$ that correspond to these poles now being equal to

$$\zeta_{\bar{m}} \triangleq \lim_{z \to z_{o,\bar{m}}} \left(1 - \frac{z}{z_{o,\bar{m}}}\right) Q(z) = -\frac{(1-\rho)\left(z_{o,\bar{m}} - 1\right)/z_{o,\bar{m}}}{1 - E'_{\bar{m}}\left(z_{o,\bar{m}}\right) T'\left(E_{\bar{m}}\left(z_{o,\bar{m}}\right)\right)} \Psi_{\bar{m}}\left(z_{o,\bar{m}}\right) \quad , \qquad \text{(IV.9b)}$$

expressions (III.32b) (with $c=1$) for the ccdf of the queue and system content remain unabatedly applicable. As before, the calculation of these quantities can be considerably facilitated by invoking approximation (III.21a) for $\Psi_{\bar{m}}(z)$.

IV.4 The unfinished work

An additional quantity that is of interest in the performance evaluation of buffers with sources that generate variable-length packets, is the so-called unfinished work, sometimes also referred to as the virtual waiting time. The unfinished work in the buffer at the beginning of a slot is defined as the amount of work in the buffer at the beginning of the slot, which is the number of slots required to empty the buffer if no more packets were to arrive during the subsequent slots. Hence, in case of single-slot packet transmission times, the unfinished work is equal to the system content. As we will see later on, the analysis of the unfinished work can be regarded as a useful intermediate step in the analysis of the packet waiting time and delay.

IV.4.1 derivation of the pgf

The number of slots that are required to process a packet whose transmission has not started yet is given by its entire transmission time, described by the pgf $T(z)$. On the other hand, at the beginning of a slot, the packet currently being transmitted (if any) still requires an amount of slots – before it is completely sent – that is equal to its residual transmission time. The unfinished work u_n at the beginning of slot n is therefore given by

$$u_n = \begin{cases} 0 & \text{, if } s_n = 0,\ \tau_n = 0 \\ \tau_n + \sum_{i=1}^{s_n - 1} t_i \text{ ,} & \text{if } s_n > 0 \end{cases} \quad ,$$

where each of the t_i's represents a complete packet transmission time, and is therefore described by the pgf $T(z)$. Consequently, the z-transform of these system equations for the unfinished work reveals that the corresponding steady-state pgf $U(z)$ satisfies

$$U(z) = T(z)^{-1} \left\{ P_s\left(T(z), z, \bar{\boldsymbol{I}}\right) + \left(T(z) - 1\right) P_s\left(0, 0, \bar{\boldsymbol{I}}\right) \right\} \quad .$$

The property that $P_s\left(0,0,\overline{\boldsymbol{I}}\right)$=Pr[$s$=0]=1-ρ can be written as

$$P_s\left(0,0,\overline{\boldsymbol{I}}\right)=\sum_{\overline{\boldsymbol{l}}} p\left(\overline{\boldsymbol{l}}\right)=\sum_{\overline{\boldsymbol{m}}}\Psi_{\overline{\boldsymbol{m}}}\left(T(z)\right) \quad ,$$

combined with the previous identity and expression (IV.4a) for $P_s(z,y,\overline{\boldsymbol{x}})$, leads to the following result for the steady-state probability generating function describing the unfinished work at the beginning of an arbitrary slot :

$$\boxed{U\left(z\right)=\left(1-\rho\right)\sum_{\overline{\boldsymbol{m}}}\frac{\left(z-1\right)E_{\overline{\boldsymbol{m}}}\left(T\left(z\right)\right)}{z-E_{\overline{\boldsymbol{m}}}\left(T\left(z\right)\right)}\Psi_{\overline{\boldsymbol{m}}}\left(T\left(z\right)\right)} \quad . \tag{IV.10a}$$

We would like to underscore that each of the denominators in the above expression also has a zero inside the unit disk $\{z\in\mathbb{C}:|z|\leq 1\}$, which leads to a an additional set of linear equations for the boundary probabilities $p\left(\overline{\boldsymbol{l}}\right)$, since the pgf $U(z)$ is bounded inside this region as well. However, if we denote the zero that stems from the denominator $z-E_{\overline{\boldsymbol{m}}}\left(T\left(z\right)\right)$ by $\tilde{z}_{0,\overline{\boldsymbol{m}}}$, then in view of the equation (IV.6b) for $z_{0,\overline{\boldsymbol{m}}}$, we may then write

$$T\left(\tilde{z}_{0,\overline{\boldsymbol{m}}}\right)=T\left(E_{\overline{\boldsymbol{m}}}\left(T\left(\tilde{z}_{0,\overline{\boldsymbol{m}}}\right)\right)\right)\Rightarrow T\left(\tilde{z}_{0,\overline{\boldsymbol{m}}}\right)=z_{0,\overline{\boldsymbol{m}}} \quad ,$$

which demonstrates that the set of linear equations for the boundary probabilities that we thus obtain,

$$\Psi_{\overline{\boldsymbol{m}}}\left(T\left(\tilde{z}_{0,\overline{\boldsymbol{m}}}\right)\right)=0 \quad ,$$

is in fact equivalent to (IV.6b), and does not impose additional conditions on these quantities.

In addition, we would also like to point out that expression (IV.10a) can also be directly derived from the results that were presented in the previous chapter. Indeed, the unfinished work at the beginning of consecutive slots satisfies the system equation

$$u_{n+1}=\left(u_n-1\right)^{+}+\sum_{i=1}^{e_n}t_i \quad , \tag{IV.10b}$$

which is equal to the system equation for the system content in a single-server queuing system, with 'minipacket' arrivals of length 1 that are generated by the pgm $\mathbf{Q}(T(z))$ instead of $\mathbf{Q}(z)$. Hence, if we take expression (III.18a) of the previous chapter for c=1, and take into consideration that the argument z of the eigenvalues and –vectors must be replaced by $T(z)$, we thus find

$$P_u(z,\bar{\boldsymbol{x}}) \triangleq \lim_{n\to\infty} E\left[z^{u_n}\bar{\boldsymbol{x}}^{\bar{\boldsymbol{a}}_{n-1}}\right] = (1-\rho)\sum_{\bar{\boldsymbol{m}}} \frac{(z-1)\left(\boldsymbol{\Lambda}(T(z))\mathbf{W}(T(z))\bar{\boldsymbol{x}}\right)^{\bar{\boldsymbol{m}}}}{z - E_{\bar{\boldsymbol{m}}}(T(z))}\Psi_{\bar{\boldsymbol{m}}}(T(z))} \quad . \quad \text{(IV.10c)}$$

Obviously, this reduces to expression (IV.10a) if we set $\bar{\boldsymbol{x}} = \bar{\boldsymbol{I}}$, which confirms the correctness of the foregoing calculations concerning the queue and system content on the one hand, and the unfinished work on the other hand.

IV.4.2 mean and variance

Closed-form formulae for the mean and variance of the unfinished work u can be computed by applying the $\boldsymbol{\mathcal{M}}[\cdot]$ and $\boldsymbol{\mathcal{V}}[\cdot]$ operators on expression (IV.10a) for $U(z)$, and adopting analogous computational rules as in Section IV.3.2. This yields the following results for the mean unfinished work

$$\mu_u = \frac{\mu_t^2\sigma_e^2\kappa_b + \mu_e\sigma_t^2}{2(1-\rho)} + \frac{\rho}{2} + \mu_t\boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{N}}}\right] \quad , \quad \text{(IV.11a)}$$

and its variance

$$\begin{aligned}\sigma_u^2 = {} & \frac{\mu_t^3\mu_{3,e}\kappa_s + 3\mu_t\sigma_t^2\sigma_e^2\kappa_b + \mu_e\mu_{3,t}}{3(1-\rho)} + \left(\frac{\mu_t^2\sigma_e^2 + \mu_e\sigma_t^2\kappa_b}{2(1-\rho)}\right)^2 + \frac{\mu_t^2\sigma_e^2\kappa_b + \mu_e\sigma_t^2}{2} \\ & + \frac{1-(1-\rho)^2}{12} + \left(\mu_t^2\boldsymbol{\mathcal{V}}\left[\Psi_{\bar{\boldsymbol{N}}}\right] + \left(2(1-\rho)\mu_t + \sigma_t^2\right)\boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{N}}}\right]\right) + 2\mu_t\sum_{\bar{\boldsymbol{m}}\in\Omega_{\bar{\boldsymbol{a}}}\setminus\bar{\boldsymbol{N}}}\frac{(1-\rho)}{1-E_{\bar{\boldsymbol{m}}}(1)}\boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{m}}}\right] ,\end{aligned} \quad \text{(IV.11b)}$$

where the expressions for $\boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{N}}}\right]$ and $\boldsymbol{\mathcal{V}}\left[\Psi_{\bar{\boldsymbol{N}}}\right]$ are of course the same as before, and where the last term in the right-hand side satisfies (III.28c) with $c=1$.

The heavy-load approximation for the mean and variance of the unfinished work yields

$$\begin{aligned}\mu_{u,\rho\to 1} &= \frac{\mu_t^2\sigma_e^2\kappa_b + \mu_e\sigma_t^2}{2(1-\rho)} \\ \sigma_{u,\rho\to 1}^2 &= \left(\frac{\mu_t^2\sigma_e^2\kappa_b + \mu_e\sigma_t^2}{2(1-\rho)}\right)^2 \quad .\end{aligned} \quad \text{(IV.11c)}$$

Using a similar argument as in the discussion concerning the heavy-load approximation of the mean and variance of the system and queue content, we observe that the numerator of this expression is likely to be dominated by the term $\mu_t^2\kappa_b\sigma_e^2$, in which case the mean and stdv of the unfinished work is (roughly) equal to respectively the mean and stdv of the queue and/or system content, multiplied by the mean packet length μ_t.

Finally, expressions such as (III.30a,b) can be invoked once again if we want to circumvent the numerical computation of the boundary probabilities.

IV.4.3 tail distribution

The tail distribution of the unfinished work ccdf can be approximated by applying the multiple-pole tail approximation technique introduced in Section III.3.6.1. The poles of $U(z)$ that ascertain this tail behaviour, for each value of $\bar{\boldsymbol{m}}$ denoted by $\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}$, will now be the solutions of the equations

$$\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}} = E_{\bar{\boldsymbol{m}}}\left(T\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}\right)\right) \equiv \prod_{j=1}^{L} \lambda_j \left(T\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}\right)\right)^{m_j} \;;\; \tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}} \in]1,\mathcal{R}[\;,\; \bar{\boldsymbol{m}} \in \tilde{\Omega}_{\mathrm{o},\bar{\boldsymbol{a}}} \quad , \tag{IV.12a}$$

where $\tilde{\Omega}_{\mathrm{o},\bar{\boldsymbol{a}}}$ contains those values of $\bar{\boldsymbol{m}}$ for which the condition

$$\lim_{\substack{z \to \mathcal{R} \\ <}} E_{\bar{\boldsymbol{m}}}(T(z))/z > 1 \quad ,$$

is satisfied. The radius of convergence $\mathcal{R}$ is now equal to the minimum of $\mathcal{R}_t$, and the value z^* defined by $T(z^*)=\mathcal{R}_q$ (if such a solution exists for $z^* < \mathcal{R}_t$). Since $T(z)$ and, generally speaking (e.g., the arguments and examples that were put forward in Section III.3.6.2), $E_{\bar{\boldsymbol{m}}}(z)$, are monotonically increasing functions in the region $]1,\mathcal{R}[$, it becomes apparent that in that case

$$E_{\bar{\boldsymbol{m}}}(T(z))/z > 1 \Leftrightarrow T(E_{\bar{\boldsymbol{m}}}(T(z)))/T(z) > 1 \Leftrightarrow T(E_{\bar{\boldsymbol{m}}}(z))/z > 1 \quad ,$$

and, therefore, $\tilde{\Omega}_{\mathrm{o},\bar{\boldsymbol{a}}} \equiv \Omega_{\mathrm{o},\bar{\boldsymbol{a}}}$. Consequently, one can deduce that the poles of $Q(z)$ and $S(z)$ on the one hand and of $U(z)$ on the other hand are related by

$$T\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}\right) = z_{\mathrm{o},\bar{\boldsymbol{m}}} \;;\; E_{\bar{\boldsymbol{m}}}\left(z_{\mathrm{o},\bar{\boldsymbol{m}}}\right) = \tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}} \;,\; \forall \bar{\boldsymbol{m}} \in \Omega_{\mathrm{o},\bar{\boldsymbol{a}}} \quad .$$

In addition, the coefficients $\zeta_{u,\bar{\boldsymbol{m}}}$ that are associated with its poles are given by

$$\zeta_{u,\bar{\boldsymbol{m}}} \triangleq \lim_{z \to z_{\mathrm{o},\bar{\boldsymbol{m}}}} \left(1 - \frac{z}{\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}}\right) U(z) = -\frac{(1-\rho)\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}} - 1\right)}{1 - T'\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}\right) E'_{\bar{\boldsymbol{m}}}\left(T\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}\right)\right)} \Psi_{\bar{\boldsymbol{m}}}\left(T\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}}\right)\right) \quad , \tag{IV.12b}$$

where the computational complexity can be considerably reduced by invoking approximation (III.21a) for $\Psi_{\bar{\boldsymbol{m}}}(z)$ when calculating these quantities. We ultimately obtain the following approximate expression for the ccdf of the unfinished work :

$$\Pr[u > T_h] \cong \sum_{\bar{m} \in \Omega_{o,\bar{a}}} \frac{\zeta_{u,\bar{m}} \tilde{z}_{o,\bar{m}}^{-T_h}}{\tilde{z}_{o,\bar{m}} - 1} \quad ,$$

that can be invoked for high enough values of the threshold T_h.

IV.5 The packet waiting time and delay

In accordance to the definitions of the previous chapters, we denote by the packet delay the number of slots that an arbitrary packet resides in the queuing system, including its transmission time, whereas the packet waiting time does not include the packet's transmission time. Both the packet waiting time and delay start at the beginning of the slot that follows on the packet's arrivals slot, as depicted in Fig. IV-2, where we have also shown some of the quantities that play a role in the analysis of the sojourn time of a packet in the system. The service discipline under consideration is FCFS for packets that do not arrive during the same slot, while packets that do arrive in the same slot will be transmitted in random order. An analysis of the waiting time and/or delay, by means of a pgf-based solution approach, has been reported in [136] and [197] in case of an i.i.d arrival process.

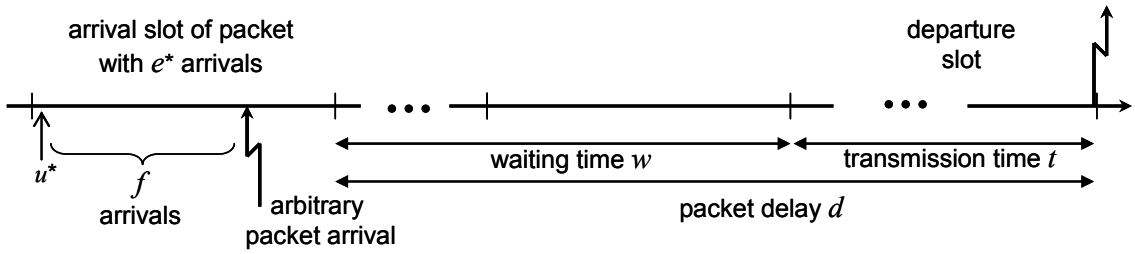


Figure IV-2 : packet waiting time and delay

IV.5.1 derivation of the pgfs $W(z)$ and $D(z)$

Consider an arbitrary tagged packet that enters the buffer during slot n. For the same reasons as in the analysis of the packet sojourn time in case of single-slot transmission times, slot n is not an arbitrary slot, but a slot where an arbitrary packet arrives (which for instance implies that at least one packet arrives). Due to the FCFS service discipline, the waiting time of a packet is determined by the amount of work in the buffer upon the tagged packet's arrival. In particular, if we denote by the drv u^* the unfinished work at the beginning of the slot during which the arbitrary packet arrives, and by f the number of packets that have arrived during the same slot as the tagged one and will be transmitted before it, then the waiting time of an arbitrarily chosen packet can be written as

$$w = (u^* - 1)^+ + \sum_{i=1}^{f} t_i \quad , \qquad \text{(IV.13a)}$$

where the random variables t_i, $1 \le i \le f$, represent the transmission times of the f packets that have arrived during its arrival slot and will be transmitted before it, which are all described by the pgf $T(z)$. Let us denote by $G(z,x)$ the steady-state joint pgf of the couple of drvs (u, e) that represent the unfinished work at the beginning, and the number of packet arrivals during, an arbitrary slot. Similarly, $H(z,x)$ represents the steady-state joint pgf of the pair (u^*,f). Then a completely analogous derivation as in Section III.4.1 shows that these two functions are related by :

$$H(z,x) = \frac{G(z,x) - G(z,1)}{\mu_e (x-1)} \quad .$$

Moreover, system equation (IV.13a) that determines the packet waiting time can be transformed into a relation between z-transforms, which in combination with the above result, leads to an expression for the steady-state pgf $W(\cdot)$ describing the waiting time of an arbitrary packet, in terms of $G(\cdot,\cdot)$

$$W(z) = \frac{1}{\mu_e z} \left\{ \frac{G(z,T(z)) - G(z,1)}{T(z)-1} + (z-1) \frac{G(0,T(z)) - G(0,1)}{T(z)-1} \right\} \quad .$$

On the other hand, since the unfinished work satisfies the system equation (IV.10b), and assuming that a stochastic equilibrium is reached, this implies that the pgf $U(z)$ can also be written as

$$U(z) = z^{-1} \{ G(z,T(z)) + (z-1) G(0,T(z)) \} \quad .$$

Inserting this relation into the previous expression for $W(z)$, and making use of $U(z)=G(z,1)$ and $U(0)=G(0,1)$, this eventually leads to

$$\boxed{W(z) = z^{-1} \frac{\mu_t (z-1)}{T(z)-1} U_0(z)} \quad , \qquad \text{(IV.13b)}$$

where, in view of $U(0)=\Pr[u=0]=\Pr[s=0]=1-\rho$, the (conditional) pgf $U_0(z)$ is defined as

$$U_0(z) \triangleq \mathbf{E}\left[z^u \,|\, u>0 \right] = \rho^{-1} (U(z) - U(0)) \qquad \text{(IV.13c)}$$

This formula for $W(z)$, which expresses the waiting time of an arbitrary packet in terms of the unfinished work at the beginning of an arbitrary slot, and which, apart from the mean arrival rate, apparently is independent of the details of the packet arrival process (type of correlation, etc…) was also derived in [259] via an alternative analysis. From this result, it is not difficult to express the performance measures concerning the packet waiting time in terms of the corresponding performance measures for the unfinished work, as shown later on.

The delay of an arbitrary packet on the other hand, is then easily derived from the relation $d=w+t$, and the observation that the transmission time of an arbitrary packet is by no means dependent on the waiting time in the queue that it has experienced (and vice versa). We may therefore write

$$\boxed{D(z)=T(z)W(z)=\frac{T(z)}{z}\frac{T_p(z-1)}{T(z)-1}U_0(z)} \quad . \tag{IV.13d}$$

It is worth noting that the factor that precedes the pgf $U_0(z)$ in expression (IV.13b) for $W(z)$ is equal to $T_{r,0}(z)^{-1}$, where $T_{r,0}(\cdot)$ represents the pgf of an arbitrary residual transmission time in a nonempty buffer, given by (IV.5a). This enables us to express performance indices such as the mean and variance of the packet waiting time and delay in terms in terms of the moments of u and τ_0 in a straightforward manner, as shown next.

IV.5.2 moments of the packet waiting time and delay

If we apply the $\mathcal{M}[\cdot]$ and $\mathcal{V}[\cdot]$ operators on both hand sides of the identity $W(z)=U_0(z)/T_{r,0}(z)$, we easily find that

$$\begin{aligned} \mu_w &= \rho^{-1}\mu_u - \mu_{\tau_0} \\ \sigma_w^2 &= \rho^{-1}\sigma_u^2 - \rho^{-2}(1-\rho)\mu_u^2 + \sigma_{\tau_0}^2 \end{aligned} \quad . \tag{IV.14a}$$

In addition, the relations for the mean and variance of the packet delay

$$\begin{aligned} \mu_d &= \mu_w + \mu_t \\ \sigma_d^2 &= \sigma_w^2 + \sigma_t^2 \end{aligned} \quad . \tag{IV.14b}$$

immediately ensue from the equation $D(z)=T(z)W(z)$.

As a powerful check on the calculations that were presented in this chapter, let us verify that the mean queue (and system) content on the one hand, and the mean packet waiting time (and delay) on the other hand satisfy Little's theorem, i.e., that the relations

$$\begin{aligned} \mu_q &= \mu_e \cdot \mu_w \\ \mu_s &= \mu_e \cdot \mu_d \end{aligned} \quad ,$$

are indeed valid. In view of the above expression for μ_w, and (IV.5b) for μ_{τ_0}, the relation $\mu_q=\mu_e\cdot\mu_w$ will hold if

$$\mu_t\mu_q + \rho\frac{\sigma_t^2}{2\mu_t} + \rho\frac{\mu_t+1}{2} = \mu_u \quad ,$$

and in view of expressions (IV.8a) and (IV.11a) for μ_q and μ_u respectively, it only takes a minor calculation to ensure oneself that this is indeed the case.

IV.5.3 tail distributions of the packet waiting time and delay

If we adopt the multiple-pole approximation technique on the pgf $W(z)=U_0(z)/T_{r,0}(z)$, it is clear that the poles of $W(z)$ that determine its tail behaviour, are those of $U_0(z)$, and hence $U(z)$, and we therefore immediately find that, for high enough values of the threshold T_h

$$\Pr[w>T_h] \cong \frac{1}{\mu_e} \sum_{\bar{m}\in\Omega_{o,\bar{a}}} \frac{\zeta_{u,\bar{m}} \tilde{z}_{o,\bar{m}}^{-T_h-1}}{T(\tilde{z}_{o,\bar{m}})-1} \quad . \tag{IV.15a}$$

In addition, in view of $D(z)=T(z)W(z)$, we readily also find that

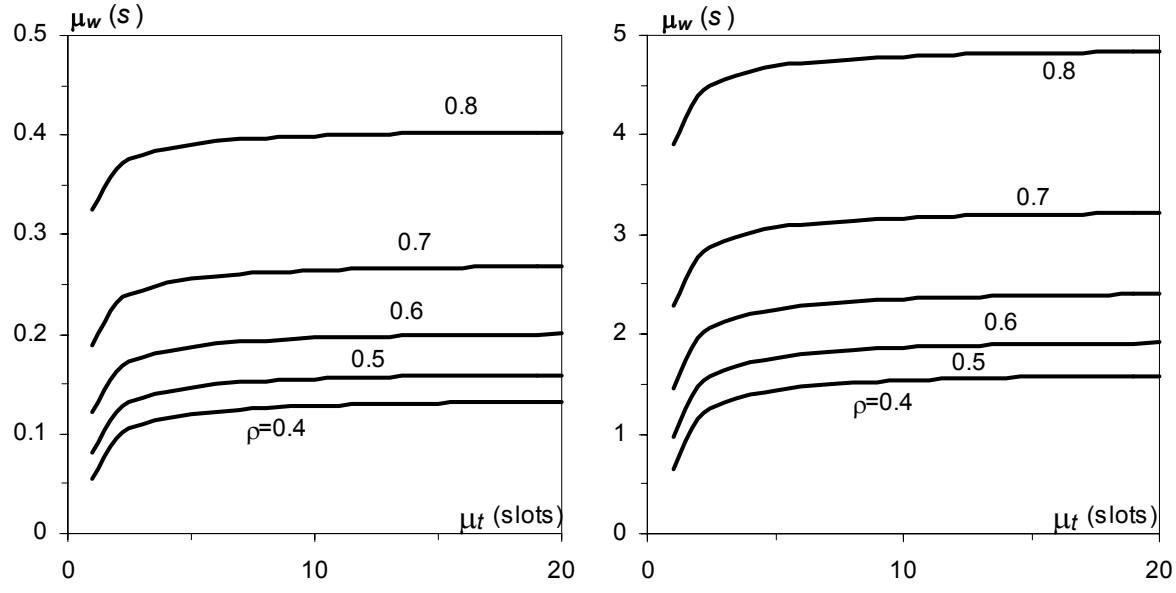
$$\Pr[d>T_h] \cong \frac{1}{\mu_e} \sum_{\bar{m}\in\Omega_{o,\bar{a}}} T(\tilde{z}_{\bar{m}}) \frac{\zeta_{u,\bar{m}} \tilde{z}_{o,\bar{m}}^{-T_h-1}}{T(\tilde{z}_{o,\bar{m}})-1} \quad , \tag{IV.15b}$$

where the poles $\tilde{z}_{o,\bar{m}}$ and the coefficients $\zeta_{u,\bar{m}}$ can be calculated – and approximated as far as the latter are concerned – as explained in Section IV.4.3.

Making use of the analysis that was presented in Section III.5 for the case of a heterogeneous D-BMAP with single-slot transmission times, in combination with the results that were deduced in this chapter up to now, it is not too difficult to extend the latter to the case of a heterogeneous D-BMAP with K traffic classes. For reasons that will become clear later on however, we need to distinguish between the case where on the one hand all packets still have the same (i.i.d.) packet transmission time distribution in common that is characterised by a single pgf $T(z)$, and on the other hand the situation where the packet transmission time of a packet of class k depends on the traffic class it belongs to, and is described by the pgf $T_k(z)$, $1\leq k\leq K$. Before resuming to the analysis of such models, let us first consider a few examples.

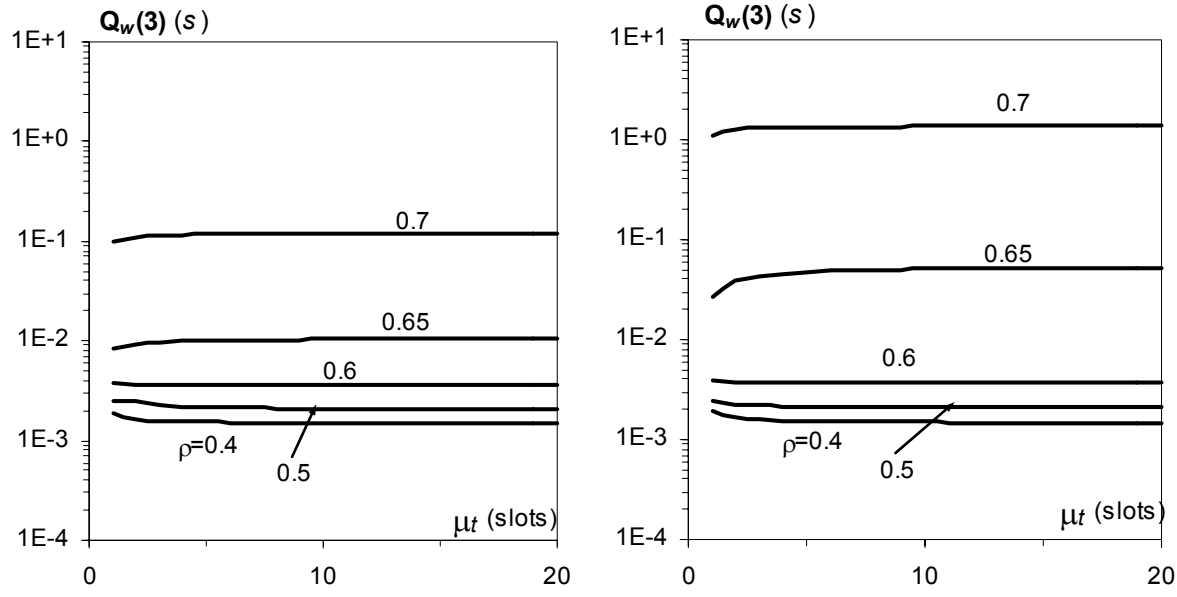
IV.5.4 some numerical examples

By means of a small set of numerical examples, based on the case study of Section III.6, we will examine the impact of the slot length, represented here by S_l, on the system performance. In that section, S_l was equal to $(8\cdot 200)/R_{link}$ for packet sizes of $200B$, and single-slot packet transmission times. We should emphasise that the slot length S_l, which is the basic time unit in our discrete-time model, is not a parameter that is an inherent characteristic of the packetised phone traffic specifications; therefore we expect that our results concerning the buffer performance are (nearly) independent of our specific choice for S_l. In order to investigate this, we will consider decreasing values of S_l in the model for phones without silence suppression, while simultaneously increasing the (constant) packet transmission times μ_t and the traffic parameters T_P, T_A, and I_a (expressed in slots) accordingly. Or, inversely, increasing the packet transmission time μ_t (expressed in slots) becomes equivalent to decreasing the slot length S_l (expressed in seconds).



(a) T_A=10s (b) T_A=120s

Figure IV-3 : mean packet waiting time versus μ_t; α=0.2, R_{link}=5Mb/s, R_{cod}=64kb/s



(a) T_A=10s (b) T_A=120s

Figure IV-4 : packet waiting time quantile versus μ_t; α=0.2, R_{link}=5Mb/s, R_{cod}=64kb/s

In Figs. IV-3a,b, for a scenario with R_{link}=5Mb/s, R_{cod}=64kb/s and an activity grade α=0.2, we have plotted the mean waiting time (expressed in seconds) versus the packet transmission time μ_t (expressed in slots), for values of the load ρ as indicated, and T_A=10s (a) and 120s (b) respectively. It becomes clear from these two figures that, although there is a slight dependence for low values of μ_t, our results are as good as independent from the slot length S_l that we choose in our model, as expected. Since we have expressed μ_w in seconds in these figures, this means that we must multiply expression (IV.14a) for this quantity by S_l. The

observation that there is in fact a small dependence on S_l, and that the mean waiting time increases as S_l decreases is explained by the following reasoning. First, note that a closer inspection of expression (III.61a) for the burst factor κ_b reveals that this quantity is completely independent of the value of S_l, since the ratio T_A/I_a remains unchanged for varying values of S_l. Secondly, the factor $S_l(\mu_t \cdot \sigma_e)^2 = N \cdot (S_l \mu_t) \cdot (\mu_t / I_a)\,(1 - 1/I_a)$ that, along with κ_b, appears in the numerator of expression (IV.11a) for μ_u (and hence, μ_w), (slightly) increases for decreasing values of S_l, and becomes equal to $N \cdot (S_l \mu_t) \cdot (\mu_t / I_a)$, a quantity that is independent of the specific value of μ_t. Hence, this accounts for the horizontal asymptote to which each of the curves in Figs. IV-3a,b evolves to. Also, observe that these curves indicate that the mean packet waiting times are proportional to T_A, since corresponding data points for $T_A = 10s$ and $T_A = 120s$ respectively, are scaled by a proportionality factor that is almost exactly equal to 12.

Similar qualitative conclusions can be drawn as well from Figs. IV-4a,b, where we have plotted the 10^{-3} quantile of the packet waiting time versus the packet length μ_t. The dependence on μ_t of these results is marginal, since these quantities almost immediately reach their limiting value for increasing μ_t. For low values of the load ρ, the quantiles exhibit a slightly decreasing tendency for increasing μ_t, while the opposite is true if the values of the load become closer to 1. For values of the quantiles that are sufficiently low, there is virtually no dependence on the specific value of T_A – an observation that was already touched upon in the discussion of Figs. III-22a,b – while a mild dependency exists for high values of the quantiles (and ρ).

Once gain, an interesting topic is the limiting behaviour that is shown in these figures for increasing μ_t. As S_l becomes (infinitesimally) small, the discrete-time queuing model converges to a continuous-time queuing model, with N arrival streams and exponentially distributed active and passive periods with mean T_A and T_P; during the active periods, packets are generated with rate $1/I_a$, while the packet transmission times are deterministic and equal to $(8 \cdot 200)/R_{link}$. This amounts to a so-called MMPP/D/1 queue (MMPP = Markov modulated Poisson process), and an algorithmic solution method for this kind of queuing model, based on a matrix-analytic approach, can for instance be found in [161].

IV.6 A heterogeneous D-BMAP with class-independent packet transmission times

Let us from now on focus on the situation where the sources that generate the packet arrivals can be subdivided into K traffic classes, where each of the classes comprises N_k sources, $1 \leq k \leq K$, and where each source of type k is characterised by means of a Markov chain with L_k states, that is fully captured by the pgm $\mathbf{Q}_k(z)$, with state space $\Omega_{\bar{a}_k}$. This is the traffic scenario that was expounded upon in Section III.5.1, and the definitions and notations that were developed there, will also be adopted in the current section. Furthermore, we now

assume that each packet that enters the output buffer requires a transmission time that is described by the pgf $T(z)$ regardless of the class that it belongs to, and that the transmission times of subsequent packets constitute a series of i.i.d. drvs. The load of such a system can then be calculated from

$$\rho = \mu_t \cdot \mu_e = \mu_t \sum_{k=1}^{K} N_k \overline{\boldsymbol{\pi}}_k^T \mathbf{Q}_k'(1) \overline{\boldsymbol{1}}_k \quad .$$

IV.6.1 the joint pgf of the state vector

For this kind of packet arrival process, the analyses that have been reported throughout this work have made clear that, in addition to the system content and the residual transmission time (if the buffer is nonempty) at the beginning of a tagged slot (say slot n+1), we will also have to keep track of the number of sources of type k in each of the possible states during the preceding slot, which shows that the set of $2+\sum_k L_k$ drvs $(s_{n+1}, \tau_{n+1}, \overline{\boldsymbol{a}}_{1,n}, \cdots, \overline{\boldsymbol{a}}_{K,n})$ will serve as the state vector for this queuing model. Therefore, our objective is to derive a closed-form expression for the steady-state joint pgf

$$P_s(z, y, \overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K) \triangleq \lim_{n\to\infty} \mathbf{E}\left[z^{s_{n+1}} y^{\tau_{n+1}} \prod_{k=1}^{K} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k,n}} \right] = \mathbf{E}\left[z^{s} y^{\tau} \prod_{k=1}^{K} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k}} \right] \quad .$$

The system equations (IV.1a-c) that were introduced and discussed in Section IV.2.1 still provide a valid description of this system's evolution in time, provided that we let e_n represent the *total* number of packet arrivals entering the buffer in slot n, i.e.,

$$e_n = \sum_{k=1}^{K} e_{k,n} \quad ,$$

where $e_{k,n}$ denotes the number of packet arrivals during slot n of type k. Then the analysis presented in Section IV.2.2 can be extended to show that the functional equation (IV.2b) is expanded into

$$\boxed{\begin{aligned} & y P_s(z, y, \overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K) = P_s(z, y, \mathbf{Q}_1(z)\overline{\boldsymbol{x}}_1, \cdots, \mathbf{Q}_K(z)\overline{\boldsymbol{x}}_K) + y(1 - T(y))\varphi(\overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K) \\ & + (yT(y) - 1)\varphi(\mathbf{Q}_1(z)\overline{\boldsymbol{x}}_1, \cdots, \mathbf{Q}_K(z)\overline{\boldsymbol{x}}_K) + y(T(y) - z)\Upsilon(z, \mathbf{Q}_1(z)\overline{\boldsymbol{x}}_1, \cdots, \mathbf{Q}_K(z)\overline{\boldsymbol{x}}_K) \end{aligned}} \quad , \quad \text{(IV.16a)}$$

with

$$\begin{cases} \varphi(\overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K) \triangleq \lim_{n\to\infty} \mathbf{E}\left[\prod_{k=1}^{K} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k,n}} \{s_{n+1} = 0, \tau_{n+1} = 0\} \right] \\ \Upsilon(z, \overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K) \triangleq \lim_{n\to\infty} \mathbf{E}\left[z^{s_{n+1}-1} \prod_{k=1}^{K} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k,n}} \{\tau_{n+1} = 1\} \right] \end{cases} ,$$

for describing the current system.

This functional equation can now be solved via similar techniques as the ones developed in Section IV.2.3, and related sections of Chapter III, which involve the decomposition of the pgms $\mathbf{Q}_k(z)$ into their corresponding eigenvalues and –vectors by means of the formula

$$\prod_{k=1}^{K}\left(\mathbf{Q}_k(z)^h\,\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{l}}_k}=\prod_{k=1}^{K}\left(\sum_{\bar{\boldsymbol{m}}_k\in\Omega_{\bar{\boldsymbol{a}}_k}}F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(z)\left(\boldsymbol{\Lambda}_k(z)^h\,\mathbf{W}_k(z)\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}\right)\quad,$$

e.g. (III.47). These derivations will not be explicitly repeated here, but can be reiterated on a step-by step basis from Section IV.2.3, by expanding the state space of the underlying Markov chain from the homogenous $\Omega_{\bar{\boldsymbol{a}}}$ into the heterogeneous $\Omega_{\bar{\boldsymbol{a}}_1,\cdots,\bar{\boldsymbol{a}}_K}$ and by taking into account relations such as the above one for the decomposition of powers of $\mathbf{Q}_k(z)$. This approach eventually yields a closed-form expression for the steady-state joint pgf of the state vector :

$$\begin{aligned}P_s\left(z,y,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K\right)=(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}&\left\{\prod_{k=1}^{K}\left(\mathbf{W}_k\left(z\right)\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}\right\}\\ \cdot&\left\{1+zy\frac{\left(E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z\right)-1\right)\left(T\left(y\right)-T\left(E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z\right)\right)\right)}{\left(y-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z\right)\right)\left(z-T\left(E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z\right)\right)\right)}\right\}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z\right)\end{aligned}\qquad\text{(IV.16b)}$$

where the sum over $\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K$ runs over all possible values which belong to the state space $\{\Omega_{\bar{\boldsymbol{a}}_1},\cdots,\Omega_{\bar{\boldsymbol{a}}_K}\}$, with

$$\begin{aligned}&\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z\right)\triangleq\sum_{\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K}\left(\prod_{k=1}^{K}F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}\left(z\right)\right)p\left(\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)\\ &p\left(\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)\triangleq\frac{1}{(1-\rho)}\Pr\left[s=0,\bar{\boldsymbol{a}}_1=\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{a}}_K=\bar{\boldsymbol{l}}_{\boldsymbol{K}}\right]\end{aligned}\quad,$$

and where $E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)$ is defined in the same way as before in (III.48b).

The boundary probabilities $p\left(\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K\right)$ that appear in the above expressions can be computed by imposing the property that the joint pgf $P_s\left(z,y,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K\right)$ is a bounded function for values $\{z,y\in\mathbb{C}:|z|\le1\wedge|y|\le1\}$, which, together with the normalisation condition $\Psi_{\bar{\boldsymbol{N}}_1\cdots\bar{\boldsymbol{N}}_K}\left(1\right)=1$, yields the set of equations

$$\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\left(z_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\right)=0\quad,\ \forall\,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K\in\tilde{\Omega}\triangleq\Omega_{\bar{\boldsymbol{a}}_1\cdots\bar{\boldsymbol{a}}_K}\setminus\left\{\bar{\boldsymbol{N}}_1\cdots\bar{\boldsymbol{N}}_K\right\}\quad,\qquad\text{(IV.17a)}$$

and where the zeroes $z_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}$ are the solutions (with modulus less than 1) of

$$z_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} = T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\left(z_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\right)\right) \ , \ \left|z_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\right| < 1 \ , \ \{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega} \quad . \tag{IV.17b}$$

Also in this case, an approximation such as (III.50) for $\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)$ can be introduced, which yields

$$\begin{aligned} \Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z) &\cong C_a \prod_{k=1}^{K} \binom{N_k}{\bar{\boldsymbol{m}}_k} \prod_{l=1}^{L_k} \left(\bar{\boldsymbol{\pi}}_k^T \mathbf{Q}_k(0)^M \, \bar{\boldsymbol{u}}_{l,k}(z)\right)^{m_{l,k}} \\ C_a^{-1} &\triangleq \prod_{k=1}^{K} \left(\bar{\boldsymbol{\pi}}_k^T \mathbf{Q}_k(0)^M \, \bar{\boldsymbol{1}}_k\right)^{N_k} \quad , \end{aligned} \tag{IV.17c}$$

where M is determined to be the smallest integer for which $C_a^{-1} \leq 1-\rho$. This kind of approach reduces all numerical computations to a minimum.

IV.6.2 the queue and system content

By inserting the arguments $\bar{\boldsymbol{x}}_k = \bar{\boldsymbol{1}}_k$, $1 \leq k \leq K$, and $y=1$ into (IV.16b), one can readily deduce the following expression for $S(z)$, the steady-state pgf of the system content

$$\boxed{S(z) = (1-\rho) \sum_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \frac{(z-1)\, T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)\right)}{z - T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)\right)} \Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)} \quad . \tag{IV.18a}$$

In addition, combination of the z-transform of the relation $q=(s-1)^+$ with this formula yields the steady-state pgf $Q(z)$, that describes the queue content at the beginning of an arbitrary slot

$$\boxed{Q(z) = (1-\rho) \sum_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \frac{(z-1)}{z - T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)\right)} \Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)} \quad . \tag{IV.18b}$$

Furthermore, if we invoke the computational rules that have been developed throughout this work when applying the $\mathcal{M}[\cdot]$ and $\mathcal{V}[\cdot]$ operators on pgfs such as the above expression for $Q(z)$, we obtain, for the average queue content

$$\begin{aligned} \mu_q &= \frac{\mu_t \kappa_b \sigma_e^2 + \mu_e^2 \sigma_t^2}{2(1-\rho)} - \frac{\rho}{2} + \mathcal{M}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K}\right] \\ \mathcal{M}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K}\right] &= \sum_{\bar{\boldsymbol{l}}_1 \cdots \bar{\boldsymbol{l}}_K} \left(\sum_{k=1}^{K} \sum_{i=1}^{L} l_{i,k} u'_{i1,k}(1)\right) p(\bar{\boldsymbol{l}}_1, \cdots, \bar{\boldsymbol{l}}_K) \end{aligned} \quad , \tag{IV.19a}$$

and for its variance

$$\sigma_q^2 = \frac{\mu_t \mu_{3,e} \kappa_s + 3\mu_e \sigma_t^2 \sigma_e^2 \kappa_b + \mu_e^3 \mu_{3,t}}{3(1-\rho)} + \left(\frac{\mu_t \sigma_e^2 \kappa_b + \mu_e^2 \sigma_t^2}{2(1-\rho)} \right)^2 - \frac{\mu_t \sigma_e^2 \kappa_b + \mu_e^2 \sigma_t^2}{2} + \frac{1-(1-\rho)^2}{12}$$

$$+ \boldsymbol{\mathcal{V}}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K} \right] + 2 \sum_{\{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega}} \frac{(1-\rho)}{1 - T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(1) \right)} \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right] \quad , \qquad \text{(IV.19b)}$$

where the burst and skew factors κ_b and κ_s of the heterogeneous D-BMAP that appear in the above expressions are defined in the same way as in Section III.5.1 (e.g. (III.44a,b)), and where

$$\boldsymbol{\mathcal{V}}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K} \right] = \sum_{\bar{\boldsymbol{l}}_1 \cdots \bar{\boldsymbol{l}}_K} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} \boldsymbol{\mathcal{V}}\left[u_{i1,k} \right] + \left(\sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} u'_{i1,k}(1) \right)^2 \right\} p(\bar{\boldsymbol{l}}_1, \cdots, \bar{\boldsymbol{l}}_K) - \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K} \right]^2$$

$$\sum_{\{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega}} \frac{1}{1 - T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(1) \right)} \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right] = \sum_{\bar{\boldsymbol{l}}_1 \cdots \bar{\boldsymbol{l}}_K} \left(\sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} \sum_{n=2}^{L_k} \frac{u'_{in,k}(1)}{1 - T\left(\lambda_{n,k}(1) \right)} \right) p(\bar{\boldsymbol{l}}_1, \cdots, \bar{\boldsymbol{l}}_K) \, .$$

Moreover, the moments of the queue and system content are currently related by

$$\begin{aligned} \mu_s &= \mu_q + \rho \\ \sigma_s^2 &= \sigma_q^2 + \kappa_b \sigma_e^2 + 2(1-\rho) \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right] \end{aligned} \quad . \qquad \text{(IV.19c)}$$

If we invoke approximation (IV.17c) for the terms in the above formulae that contain the boundary probabilities, then we obtain approximations that are equivalent to (III.53c), where the last of these three equations must presently be replaced by

$$\sum_{\{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega}} \frac{1}{1 - T\left(E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(1) \right)} \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right] \cong \sum_{k=1}^{K} N_k \sum_{n=2}^{L_k} \frac{\bar{\boldsymbol{\pi}}_k^T \mathbf{Q}_k(0)^M \, \bar{\boldsymbol{u}}'_{n,k}(1)}{\left(1 - T\left(\lambda_{n,k}(1) \right) \right) \left(\bar{\boldsymbol{\pi}}_k^T \mathbf{Q}_k(0)^M \, \bar{\boldsymbol{l}}_k \right)} \quad .$$

Also observe that the heavy-load limits of the mean and variance of the queue and system content are given by the same formal expressions as in the homogeneous case, e.g. expressions (IV.8d).

Finally, as far as the tail distribution of the queue and system content is concerned, we can again adopt a multiple pole approximation technique; which basically means that we extend the state space from $\Omega_{\bar{\boldsymbol{a}}}$ to $\Omega_{\bar{\boldsymbol{a}}_1 \cdots \bar{\boldsymbol{a}}_K}$ for the results that were reported in Section IV.3.3 (hence, the column vector $\bar{\boldsymbol{m}}$ is replaced by the set of vectors $\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K$). Due to the close resemblance of the formulae that one thus obtains, they will not be explicitly recapitulated at this instance.

IV.6.3 the unfinished work and packet sojourn times

From the discussion in Section III.5.3, it became apparent that in the case of a heterogeneous packet arrival process, we need to specify the order by which the packets that arrive during the same slot and belong to a different traffic class will be transmitted. This has induced us to consider the so-called AO and FOC ordering schemes, whereby the former corresponds to a random ordering over all traffic classes, while the latter assigns priority to packets of traffic class k over packets of traffic class k' that enter the buffer during the same slot, if $k<k$'. We normally start from the premise that packets that belong to the same class are 'indistinguishable', i.e., treated in an equal manner, and therefore transmitted in random order if they have their arrival slot in common.

Hence, as demonstrated in Section III.5.3, the following step would be to produce an analysis of the unfinished work for the current traffic scenario, and from there on, construct the relevant equations that express the packet waiting time and delay in terms of this quantity. As it turns out however, the results that we would thus obtain are, in effect, merely a special case of those that are generated by analysing the model that is considered in the next part, which is why we choose not to explicitly present them at this point. Whenever appropriate, we will indicate how the results of the next section can be reduced to the current traffic scenario.

IV.7 A heterogeneous D-BMAP with class-dependent packet transmission times

In this section, we consider an identical heterogeneous D-BMAP as in the previous one, with the additional modelling assumption that any packet that belongs to traffic class k has a transmission time that is described by the pgf $T_k(z)$ with mean μ_{t_k}, $1\leq k\leq K$, and is therefore class-dependent. Consequently, the transmission times of consecutive packets are still assumed to constitute a series of independent random variables, but that are no longer identically distributed. In the present case, the system load is given by

$$\rho = \sum_{k=1}^{K} \mu_{t_k} \cdot \mu_{e_k} = \sum_{k=1}^{K} \mu_{t_k} \cdot N_k \overline{\boldsymbol{\pi}}_k \mathbf{Q}'(1) \overline{\boldsymbol{I}}_k \quad .$$

Unfortunately, by doing so, we introduce an insuperable difficulty in our model, at least as far as the analysis of the queue and system content is concerned. Indeed, in view of the system equations (IV.1a-c) that govern such a system, each time that the transmission of a packet is terminated and the queue is not empty (i.e., for $\tau_n=1$ and $s_n>0$), we need to be able to determine the traffic class that the next packet belongs to, in order to be able to fix the value of τ_{n+1}. Therefore, for this traffic scenario, it does not suffice to keep track of the total system content, or even the system content of each traffic class, in our state vector : we also need information on the precise order in which packets of different classes are stored in the queue.

Only then can we determine the type of the next packet that is to be transmitted after the previous one has ended. It should be clear that such a system description will produce a state vector that is incredibly difficult to keep track of at successive slot boundaries, and the corresponding (steady-state) analysis of such a vector is close to being unfeasible to the best of our knowledge. Obviously, for $T_k(z) \equiv T(z)$, $1 \le k \le K$, this difficulty does not occur, and the current model (and corresponding results) is reduced to the one that was analysed in Section IV.6. This also explains why a separate analysis for the queue and system content was provided in that section, since such an analysis cannot be carried out for the present traffic scenario.

Nevertheless, we are still able to analyse this system through the amount of work that it contains at the start of consecutive slot boundaries, as shown next.

IV.7.1 the unfinished work

If we regard a packet of any type and length t, as an entity that brings an amount of work into the buffer that is equal to t (or, equivalently, a packet of length t consists of t minipackets of length 1), then the unfinished work u_n, which represents the amount of work (or, equivalently, the number of minipackets) in the buffer at the beginning of slot n, will at subsequent slot boundaries satisfy the system equation

$$u_{n+1} = (u_n - 1)^+ + \sum_{k=1}^{K} \sum_{i=1}^{e_{k,n}} t_{i,k} \quad ,$$

where $e_{n,k}$ represents the number of packet arrivals of type k during slot n, and where $t_{i,k}$ (with pgf $T_k(z)$) denotes the length, or transmission time, of the i-th packet of type k that enters the output buffer during the tagged slot. Since each type-k packet arrival generates an amount of work describe by the pgf $T_k(z)$, then, in view of the definition of the pgm $\mathbf{Q}_k(\cdot)$ in Section III.5.1, it is not difficult to deduce that the matrix $\mathbf{Q}_k(T_k(z))$ fully describes the Markovian arrival process of the amount of work of type k that joins the buffer during consecutive slots, and this matrix will play an identical role as the pgm $\mathbf{Q}_k(z)$ in the analysis of the output buffer with a heterogeneous D-BMAP presented in Section III.5.

Thus, the above system equation for the unfinished work is in fact equivalent to that of the system content in case of single-slot transmission times, e.g. (III.45), whereby arrivals (of minipackets instead of packets) are now generated by the pgm $\mathbf{Q}_k(T_k(z))$. Therefore, if we define the steady-state joint pgf $P_u(z, \overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K)$ of the unfinished work at the beginning of a random slot, and the state of the heterogeneous D-BMAP during the foregoing slot,

$$P_u(z, \overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_K) \triangleq \lim_{n \to \infty} \mathbf{E}\left[z^{u_{n+1}} \prod_{k=1}^{K} \overline{\boldsymbol{x}}_k^{\,\overline{\boldsymbol{a}}_{k,n}} \right] \quad ,$$

we can adopt the result (III.48a) for this joint pgf, keeping in mind that we must replace the argument z of the pgm $\mathbf{Q}_k(z)$, and its eigenvalues and –vectors, by $T_k(z)$. Following this approach, we obtain

$$P_u(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)=(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)\prod_{k=1}^{K}\left(\boldsymbol{\Lambda}_k(T_k(z))\mathbf{W}_k(T_k(z))\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}}{z-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z), \quad \text{(IV.20a)}$$

where we now have that

$$\begin{aligned}
&\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)\triangleq\sum_{\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K}\left(\prod_{k=1}^{K}F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(T_k(z))\right)p(\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K)\\
&p(\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K)\triangleq\frac{1}{(1-\rho)}\Pr\left[u=0,\bar{\boldsymbol{a}}_1=\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{a}}_K=\bar{\boldsymbol{l}}_K\right] \quad . \quad \text{(IV.20b)}\\
&E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)\triangleq\prod_{k=1}^{K}\left(\boldsymbol{\Lambda}(T_k(z))\bar{\boldsymbol{l}}_k\right)^{\bar{\boldsymbol{m}}_k}\equiv\prod_{k=1}^{K}\prod_{i=1}^{L_k}\lambda_{i,k}(T_k(z))^{m_{i,k}}
\end{aligned}$$

Another quantity, that will prove to be useful in the analysis of the packet waiting time, is the drv u_q, defined as

$$u_q\triangleq(u-1)^+ \quad ,$$

which is reminiscent of the relation between the queue and system content in case of single-slot transmission times, and can be interpreted as the amount of work that resides in the system at the beginning of an arbitrary slot, not including the work unit being processed (if any) during this slot. Making use of the relation that

$$\begin{aligned}
P_{u_q}(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)&\triangleq\lim_{n\to\infty}\mathbf{E}\left[z^{(u_{n+1}-1)^+}\prod_{k=1}^{K}\bar{\boldsymbol{x}}_k^{\,\bar{\boldsymbol{a}}_{k,n}}\right]=z^{-1}\Big\{\ P_u(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)\\
&+(1-\rho)(z-1)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\sum_{\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K}\prod_{k=1}^{K}\left\{\left(\mathbf{W}_k(T_k(z))\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}F^{(k)}_{\bar{\boldsymbol{l}}_k\bar{\boldsymbol{m}}_k}(T_k(z))\right\}p(\bar{\boldsymbol{l}}_1,\cdots,\bar{\boldsymbol{l}}_K)\Big\},
\end{aligned}$$

we obtain

$$P_{u_q}(z,\bar{\boldsymbol{x}}_1,\cdots,\bar{\boldsymbol{x}}_K)=(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)\prod_{k=1}^{K}\left(\mathbf{W}_k(T_k(z))\bar{\boldsymbol{x}}_k\right)^{\bar{\boldsymbol{m}}_k}}{z-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z) \quad . \quad \text{(IV.20c)}$$

Consequently, substitution of the arguments $\bar{\boldsymbol{x}}_k=\bar{\boldsymbol{I}}_k$, $1\le k\le K$, into (IV.20a,c), produces an expression for the steady-state pgfs $U_q(z)$ and $U(z)$, of the drvs u_q and u respectively

$$\boxed{U_q(z) = (1-\rho) \sum_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \frac{(z-1)}{z - E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)} \Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)} \quad . \tag{IV.21a}$$

$$\boxed{U(z) = (1-\rho) \sum_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \frac{(z-1) E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)}{z - E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)} \Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)} \quad . \tag{IV.21b}$$

The boundary probabilities $p(\bar{\boldsymbol{l}}_1, \cdots, \bar{\boldsymbol{l}}_K)$ that appear in the above expressions can once more be computed by expressing that the joint pgfs $P_u(z, \bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_K)$ and $P_{u_q}(z, \bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_K)$, or equivalently, the pgfs $U(z)$ and $U_q(z)$, are bounded functions for values of $\{z \in \mathbb{C} : |z| \leq 1\}$, which, together with the normalisation condition $\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K}(1) = 1$, yields the set of equations

$$\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\left(\tilde{z}_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\right) = 0 \quad , \ \forall \{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega} \quad , \tag{IV.22a}$$

where $\tilde{\Omega}$ is defined as in (IV.17a), and where the zeroes $\tilde{z}_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}$ are now the solutions (with modulus less than 1) of

$$\tilde{z}_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} = E_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\left(\tilde{z}_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\right) , \ \left|\tilde{z}_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}\right| < 1 , \{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega} \quad . \tag{IV.22b}$$

Also in this case, an approximation such as (IV.17c) for $\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z)$ can be introduced, which leads to

$$\Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}(z) \cong C_a \prod_{k=1}^{K} \binom{N_k}{\bar{\boldsymbol{m}}_k} \prod_{l=1}^{L_k} \left(\bar{\boldsymbol{\pi}}_k^T \mathbf{Q}_k(0)^M \, \bar{\boldsymbol{u}}_{l,k}(T_k(z))\right)^{m_{l,k}} \quad , \tag{IV.22c}$$

where the normalisation constant C_a satisfies the same equation as in (IV.17c), and where the parameter M is determined in an identical manner.

In addition, taking the appropriate derivatives with respect to z for $z=1$ of (IV.20c), while taking into account related results that have been derived throughout this work, yields an expression for the mean value of u_q

$$\mu_{u_q} = \sum_{k=1}^{K} \frac{\mu_{t_k}^2 \cdot \sigma_{e_k}^2 \kappa_{b,k} + \mu_{e_k} \sigma_{t_k}^2}{2(1-\rho)} - \frac{\rho}{2} + \mathcal{M}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K}\right] \quad , \tag{IV.23a}$$

$$\mathcal{M}\left[\Psi_{\bar{\boldsymbol{N}}_1 \cdots \bar{\boldsymbol{N}}_K}\right] = \sum_{\bar{\boldsymbol{l}}_1 \cdots \bar{\boldsymbol{l}}_K} \left(\sum_{k=1}^{K} \mu_{t_k} \cdot \sum_{i=1}^{L} l_{i,k} u'_{i1,k}(1) \right) p(\bar{\boldsymbol{l}}_1, \cdots, \bar{\boldsymbol{l}}_K)$$

and the variance of this quantity

$$\sigma_{u_q}^2 = \sum_{k=1}^{K} \frac{\mu_{t_k}^3 \mu_{3,e_k} \kappa_{s,k} + 3\mu_{t_k} \cdot \sigma_{t_k}^2 \sigma_{e_k}^2 \kappa_{b,k} + \mu_{e_k} \mu_{3,t_k}}{3(1-\rho)} + \left(\sum_{k=1}^{K} \frac{\mu_{t_k}^2 \sigma_{e_k}^2 \kappa_{b,k} + \mu_{e_k} \sigma_{t_k}^2}{2(1-\rho)} \right)^2$$

$$- \sum_{k=1}^{K} \frac{\mu_{t_k}^2 \kappa_{b,k} \sigma_{e_k}^2 + \mu_{e_k} \sigma_{t_k}^2}{2(1-\rho)} + \frac{1-(1-\rho)^2}{12} + \boldsymbol{v}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] \qquad \text{(IV.23b)}$$

$$+2 \sum_{\{\bar{m}_1 \cdots \bar{m}_K\} \in \tilde{\Omega}} \frac{(1-\rho)}{1 - E_{\bar{m}_1 \cdots \bar{m}_K}(1)} \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{m}_1 \cdots \bar{m}_K}\right],$$

where

$$\boldsymbol{v}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] \triangleq \sum_{\bar{l}_1 \cdots \bar{l}_K} \sum_{k=1}^{K} \left\{ \sum_{i=1}^{L} l_i \left(\mu_{t_k}^2 \boldsymbol{v}\left[u_{i1,k}\right] + u'_{i1,k}(1) \sigma_{t_k}^2 \right) + \left(\sum_{i=1}^{L} l_i \mu_{t_k} u'_{i1}(1) \right)^2 \right\} p(\bar{l}_1, \cdots, \bar{l}_K)$$

$$- \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right]^2$$

$$\sum_{\{\bar{m}_1 \cdots \bar{m}_K\} \in \tilde{\Omega}} \frac{1}{1 - E_{\bar{m}_1 \cdots \bar{m}_K}(1)} \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{m}_1 \cdots \bar{m}_K}\right] = \sum_{\bar{l}_1 \cdots \bar{l}_K} \left(\sum_{k=1}^{K} \sum_{i=1}^{L_k} l_{i,k} \sum_{n=2}^{L_k} \frac{\mu_{t_k} u'_{in,k}(1)}{1 - \lambda_{n,k}(1)} \right) p(\bar{l}_1, \cdots, \bar{l}_K) \quad .$$

Furthermore, the moments of the unfinished work u can be expressed in terms of the moments of u_q by making use of the relations

$$\mu_{u_q} = \mu_u + \rho$$

$$\sigma_{u_q}^2 = \sigma_u^2 + \sum_{k=1}^{K} \left(\mu_{t_k}^2 \kappa_{b,k} \sigma_{e_k}^2 + \mu_{e_k} \sigma_{t_k}^2 \right) + 2(1-\rho) \boldsymbol{\mathcal{M}}\left[\Psi_{\bar{N}_1 \cdots \bar{N}_K}\right] \quad . \qquad \text{(IV.23c)}$$

Obviously, these results can be reduced to that of a state-independent packet transmission time scenario, by inserting

$$\mu_{t_k} \equiv \mu_t \ ; \ \sigma_{t_k}^2 \equiv \sigma_t^2 \ ; \ \mu_{3,t_k} \equiv \mu_{3,t} \quad ,$$

into these relations. Furthermore, in a similar way as before, we can establish approximations for the terms that contain the boundary probabilities in these expressions for the moments of the queue content, by taking the appropriate derivatives with respect to z for z=1 of the approximate formula (IV.22c) for $\Psi_{\bar{m}_1 \cdots \bar{m}_K}(z)$, as demonstrated before on several occasions.

As far as the heavy-load limit of the mean and variance of the unfinished work is concerned, we now obtain

$$\mu_{u_q,\rho\to 1} = \mu_{u,\rho\to 1} = \sum_{k=1}^{K} \frac{\mu_{t_k}^2 \sigma_{e_k}^2 \kappa_{b,k} + \mu_{e_k} \sigma_{t_k}^2}{2(1-\rho)}$$
$$\sigma_{u_q,\rho\to 1}^2 = \sigma_{u,\rho\to 1}^2 = \left(\sum_{k=1}^{K} \frac{\mu_{t_k}^2 \sigma_{e_k}^2 \kappa_{b,k} + \mu_{e_k} \sigma_{t_k}^2}{2(1-\rho)} \right)^2 \quad . \tag{IV.23c}$$

Apparently, the coefficient of variation of these drvs also evolves to 1 for $\rho\to 1$, as in the previous traffic scenarios investigated in this and the foregoing chapters.

To conclude, a multiple-pole tail approximation for the unfinished work ccdf can be deduced by applying similar approximation techniques as introduced in Section III.3.6.1. The poles of $U_q(z)$ and $U(z)$ that determine its tail behaviour, for each value of $\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K$ denoted by $\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}$, will now be the (smallest real) solutions of the equations

$$\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} = \prod_{k=1}^{K} \prod_{j=1}^{L} \lambda_{j,k} \left(T_k \left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right) \right)^{m_{j,k}} \; ; \begin{cases} \tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \in]1, \mathcal{R}[\\ \{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \tilde{\Omega}_{\mathrm{o},\bar{\boldsymbol{a}}_1 \cdots \bar{\boldsymbol{a}}_K} \end{cases} , \tag{IV.24a}$$

where $\tilde{\Omega}_{\mathrm{o},\bar{\boldsymbol{a}}_1,\cdots,\bar{\boldsymbol{a}}_K}$ is the set that contains those values of $\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K$ for which the inequality

$$\lim_{z \to \mathcal{R}} z^{-1} \prod_{k=1}^{K} \prod_{j=1}^{L} \lambda_{j,k} \left(T_k(z) \right)^{m_{j,k}} > 1 \quad ,$$

is fulfilled, and where $\mathcal{R}$ is defined as in (IV.12a). The coefficients $\zeta_{u_q,\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}$ that correspond to these poles satisfy

$$\begin{aligned} \zeta_{u_q,\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} &\triangleq \lim_{z \to \tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}} \left(1 - \frac{z}{\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}} \right) U_q(z) \\ &= -\frac{(1-\rho)\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} - 1 \right)}{\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \left(1 - E'_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right) \right)} \Psi_{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \right) \end{aligned} , \tag{IV.24b}$$

where the computational complexity can be considerably reduced by adopting approximation (IV.22c) for $\Psi_{\bar{\boldsymbol{m}}_1,\cdots,\bar{\boldsymbol{m}}_K}(z)$ when calculating these quantities. We ultimately obtain the following approximate expression for the ccdf of the drvs u and u_q that represent the amount of work at the beginning of an arbitrary slot, respectively with and without the work unit that is being processed :

$$\Pr[u_q > T_h] = \Pr[u > T_h + 1] \cong \sum_{\{\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K\} \in \Omega_{\mathrm{o},\bar{\boldsymbol{a}}_1,\cdots,\bar{\boldsymbol{a}}_K}} \frac{\zeta_{u_q,\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} \cdot \tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K}^{-T_h}}{\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1 \cdots \bar{\boldsymbol{m}}_K} - 1} \quad , \tag{IV.24c}$$

that is usable for high enough values of the threshold T_h.

IV.7.2 packet waiting time and delay

The results of Section III.5.3, albeit for the single-server case c=1, can now be extended to include the current scenario of class-dependent packet transmission times with a general distribution.

First, consider a packet of type k that enters the buffer under either the AO of FOC ordering discipline, which were defined and discussed in Section III.5.3. Adopting similar notations, with the addition of the subscript k to highlight the dependency on the traffic class k that the tagged packet belongs to, then the waiting time of the tagged packet satisfies

$$w_k = u^*_{q,k} + \sum_{r=1}^{K} \sum_{i=1}^{f_{k,r}} t_{i,r} \quad , \tag{IV.25}$$

where we let $u^*_{q,k}$ represent the unfinished work in the output buffer at the start of the tagged packet's arrival slot (not including the work unit that is being processed during this slot, if any), e^*_r the total number of packet arrivals of type r during the tagged slot, $f_{k,r}$ the number of packet arrivals of type r that have arrived during the same slot, and will be transmitted before it, and $t_{i,r}$, with pgf $T_r(z)$, their respective transmission times. The drvs $u^*_{q,k}$ and $f_{k,r}$ are dependent on the AO of FOC ordering process under consideration, as will become clear in the next two sections.

IV.7.2.1 arbitrary order

Consider an arbitrary packet that belongs to the k-th traffic class and enters the buffer under the AO paradigm. The total number of packets entering in the same slot that will be transmitted before the tagged one is still a uniformly distributed drv that is bounded by 0 and the total number of packet arrivals minus 1, regardless of the queue content at the start of the tagged packet's arrival slot. Hence, we may write

$$\Pr\left[\sum_{k=1}^{K} f_{k,r} = i \,\middle|\, u^*_{q,k} = j, e^*_1 = l_1, \cdots, e^*_K = l_K\right] = \frac{1}{l} \;\; ; \; 0 \le i < l \;\; , \;\; l_k > 0 \;\; , \; l \triangleq \sum_{r=1}^{K} l_r \quad ,$$

Furthermore, the system equation (IV.25) shows that it does not suffice to know the sum of the $f_{k,r}$'s, due to the differentiation in the amount of work that is generated by packets of different types. It is therefore necessary to know the individual values of the drvs $f_{k,r}$, $1 \le r \le K$, and we can show that

$$\Pr\left[f_{k,1} = i_1, \cdots, f_{k,K} = i_K \,\middle|\, u^*_{q,k} = j, \sum_{r=1}^{K} f_{k,r} = i, e^*_1 = l_1, \cdots, e^*_K = l_K\right] = \frac{l(l_k - i_k)}{l_k(l-i)} \binom{l}{i}^{-1} \prod_{r=1}^{K} \binom{l_r}{i_r} \; ; \; i \triangleq \sum_{r=1}^{K} i_r \;\; ,$$

which corresponds to the fraction of combinations by which the remaining $l_1,\cdots,l_k$-1,$\cdots,l_K$ arrivals can be rearranged in such a way that $i_1,\cdots,i_K$ of them will be transmitted before the tagged one.

On the other hand, since the tagged packet has been arbitrarily picked among all type-k packet arrivals, we can use the same arguments as in Section III.5.3 to show that

$$\Pr\left[u^*_{q,k}=j,e_1^*=l_1,\cdots,e_K^*=l_K\right]=\frac{l_k}{\mu_{e_k}}\Pr\left[u_q=j,e_1=l_1,\cdots,e_K=l_K\right] \quad , \tag{IV.26}$$

where μ_{e_k} represents the average number of type-k packet arrivals per slot, as usual. Hence, the combination of the latter three expressions with the system equation (IV.25)) results in the following expression for the pgf of w_k

$$W_k(z)\triangleq\mathbf{E}\left[z^{w_k}\right]=\mathbf{E}\left[z^{u^*_{q,k}}\prod_{r=1}^{K}T_r(z)^{f_{k,r}}\right]$$

$$=\frac{1}{\mu_{e_k}}\sum_{j=0}^{\infty}z^j\sum_{l_1=0}^{\infty}\sum_{i_1=0}^{l_1}\cdots\sum_{l_K=0}^{\infty}\sum_{i_K=0}^{l_K}\frac{(l_k-i_k)}{(l-i)}\binom{l}{i}^{-1}\left\{\prod_{r=1}^{K}\binom{l_r}{i_r}T_r(z)^{i_r}\right\}\cdot\Pr\left[u_q=j,e_1=l_1,\cdots,e_K=l_K\right].$$

Observe that this expression for the pgf $W_k(z)$ is indeed normalised, due to the property that

$$\sum_{i_1=0}^{l_1}\cdots\sum_{i_K=0}^{l_K}\frac{(l_k-i_k)}{(l-i)}\binom{l}{i}^{-1}\prod_{r=1}^{K}\binom{l_r}{i_r}=l_k \;\; ; \; i\triangleq\sum_{r=1}^{K}i_r\;\wedge\; l\triangleq\sum_{r=1}^{K}l_r \quad .$$

Also, it is not difficult to verify that this indeed reduces to the homogeneous result for K=1. Unfortunately, the combinatorial factors that appear in the sum of the penultimate equation, preclude a further simplification of this result into a closed-form expression for the joint pgf of the drvs u_q and $e_1,\cdots,e_K$. Therefore, we also consider the alternative FOC scheduling mechanism in the following section, which does allow us to derive closed-form expressions for the relevant performance characteristics, such as the mean value and/or the asymptotic tail behaviour of the waiting time of a class-k packet. As explained before, we do not expect that major differences will occur between the performance indices in case of AO ordering on the one hand and FOC ordering the other hand, in particular if the heterogeneous D-BMAP represents a bursty packet arrival process.

IV.7.2.2 fixed-order-by-class

We once more focus on an arbitrary packet of type k that enters the buffer during a slot, under the FOC ordering scheme, meaning that all packets of type r, $r<k$, that arrive during the same slot will be positioned in the queue before all type-k packet arrivals, while all packets of type r, $r>k$, that arrive during the course of this slot will be placed in the queue behind the type-k

packets. As indicated before, among the type-k packets that enter the system during the tagged packet's arrival slot, an arbitrary ordering scheme remains valid. This implies that, as in the case of single-slot transmission times, $f_{k,r}$ presently will be equal to e_r if $r<k$ and 0 if $r>k$, while $f_{k,k}$ is a uniformly distributed drv which is bounded by 0 and the total number of type-k packet arrivals minus 1. Hence, for this ordering scenario we may write

$$\mathbf{E}\left[\prod_{r=1}^{K} x_r{}^{f_{k,r}}\right] = \mathbf{E}\left[\prod_{r=1}^{k-1} x_r{}^{e_r} \cdot \frac{x_k{}^{e_k}-1}{e_k(x_k-1)}\right] \quad ,$$

and with the aid of (IV.25) and (IV.26), we can also show that

$$W_k(z) = \mathbf{E}\left[z^{u^*_{q,k}} \prod_{r=1}^{K} T_r(z)^{f_{k,r}}\right] = \frac{1}{\mu_{e_k}} \mathbf{E}\left[z^{u_q} \prod_{r=1}^{k-1} T_r(z)^{e_r} \cdot \frac{T_k(z)^{e_k}-1}{(T_k(z)-1)}\right] \quad . \tag{IV.27}$$

By taking the appropriate derivatives with respect to z for z=1 of both hand sides of this expression for the steady-state pgf $W_k(z)$, we can deduce that the mean and variance of the waiting time of a type-k packet can be computed from

$$\begin{aligned}
\mu_{e_k} \cdot \mu_{w_k} &= \mathbf{E}\left[u_q \cdot e_k\right] + \sum_{r=1}^{k-1} \rho_r + \frac{\mu_{t_k}}{2} \mathbf{E}[e_k(e_k-1)] \\
\mu_{e_k} \cdot \sigma^2_{w_k} &= \mathbf{E}\left[\left(u_q + \sum_{r=1}^{k-1} \mu_{t_r} e_r\right)^2 \cdot e_k\right] + \mu_{t_k} \mathbf{E}\left[\left(u_q + \sum_{r=1}^{k-1} \mu_{t_r} e_r\right) \cdot e_k(e_k-1)\right] \\
&\quad + \frac{\mu^2_{t_k}}{3} \mathbf{E}[e_k(e_k-1)(e_k-2)] + \frac{\mathbf{E}\left[t_k^2\right]}{2} \mathbf{E}[e_k(e_k-1)] - \mu_{e_k} \cdot \mu^2_{w_k} \quad .
\end{aligned} \tag{IV.28}$$

Again, it is interesting to note that these two expressions do not depend on the specifics of the (correlated) packet arrival process. In order to calculate the remaining expected values that contain the drv u_q in the right-hand side of these formulae, we need to establish an expression for the joint pgf

$$H_k(z, y_1, \cdots, y_k) \triangleq \mathbf{E}\left[z^{u_q} \prod_{r=1}^{k} y_r{}^{e_r}\right] \quad .$$

Adopting a similar calculus as in Section III.5.3.2 for the derivation of this kind of function, we can deduce that

$$H_k(z, y_1, \cdots, y_k) = P_{u_q}\left(z, \mathbf{Q}_1(y_1)\overline{\boldsymbol{I}}_1, \cdots, \mathbf{Q}_k(y_k)\overline{\boldsymbol{I}}_k, \overline{\boldsymbol{I}}_{k+1}, \cdots, \overline{\boldsymbol{I}}_K\right) \quad ,$$

which, in view of (IV.20c), becomes

$$H_k(z,y_1,\cdots,y_k)=(1-\rho)\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{(z-1)\Theta_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,k}(z,y_1,\cdots,y_k)}{z-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)$$

$$\Theta_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,k}(z,y_1,\cdots,y_k)\triangleq\prod_{r=1}^{k}\left(\mathbf{W}_k(T_k(z))\mathbf{Q}_k(y_k)\bar{\boldsymbol{I}}_k\right)^{\bar{\boldsymbol{m}}_k} \quad .$$

The remaining unknown quantities that occur in the right-hand side of expression (IV.28) for the mean and variance of the type-k packet waiting time can then be computed by taking the appropriate derivatives of the above joint pgf. For instance, by taking the first-order derivative with respect to z and y_k for $z=y_1=\cdots=y_k=1$, we are able to show that, in an analogous way as the calculations that led to equation (III.56b)

$$\mathbf{E}\left[u_q\cdot e_k\right]=\mu_{e_k}\cdot\mu_{u_q}+\mu_{w_k}\frac{(\kappa_{b,k}-1)\sigma_{e_k}^2}{2}+(1-\rho)\mu_{w_k}\sum_{\bar{\boldsymbol{l}}_1\cdots\bar{\boldsymbol{l}}_K}\sum_{i=1}^{L_k}l_{i,k}u'_{i1,k}(1)\,p(\boldsymbol{l}_1,\cdots,\bar{\boldsymbol{l}}_K) \quad .$$

Expressions for the expected values that remain to be determined in formula (IV.28) for the variance of the type-k waiting time can be derived in a similar manner, although the calculations become somewhat more tedious.

Finally, for the purpose of establishing a multiple-pole approximation for the ccdf of w_k, we first observe that equation (IV.27) shows that $W_k(z)$ can be calculated from

$$W_k(z)=\frac{H_k(z,T_1(z),\cdots,T_k(z))-H_k(z,T_1(z),\cdots,T_{k-1}(z),1)}{\mu_{e_k}(T_k(z)-1)}$$

$$=\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{c(1-\rho)(z-1)\tilde{\Theta}_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,k}(z)}{z^c-E_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z)}\Psi_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}(z) \quad ,$$

with

$$\tilde{\Theta}_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,k}(z)\triangleq\frac{1}{\mu_{e_k}(T_k(z)-1)}\left(\prod_{i=1}^{L_k}\lambda_{i,k}(T_k(z))^{m_{i,k}}-1\right)\left(\prod_{r=1}^{k-1}\prod_{j=1}^{L_r}\lambda_{j,r}(T_r(z))^{m_{j,r}}\right) \quad ,$$

where we have made use of the property that $\mathbf{W}_r(T_r(z))\mathbf{Q}_r(T_r(z))\bar{\boldsymbol{I}}_r=\boldsymbol{\Lambda}_r(T_r(z))\bar{\boldsymbol{I}}_r$, for all $1\le r\le K$. The multiple-pole approximation for calculating the ccdf of w_k thus yields

$$\Pr[w_k>T_h]\cong\sum_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\frac{\tilde{\Theta}_{\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K,k}\left(\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}\right)\zeta_{u_q,\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}}{\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}-1}\tilde{z}_{\mathrm{o},\bar{\boldsymbol{m}}_1\cdots\bar{\boldsymbol{m}}_K}^{-T_h} \quad . \qquad \text{(IV.29)}$$

Consequently, contrary to the AO ordering scenario, these results reveal that the moments, as well as the tail approximation of the type-k waiting time ccdf in case of FOC ordering, can indeed be calculated by an expedient application of the computational techniques that have been introduced throughout this work.

Chapter V

Conclusions and main results

In this work we have presented the analysis of a number of queuing models, with the purpose of deriving (semi-)analytic closed-form expressions for the main performance indices related to the buffer content and packet sojourn time, that provide useful tools for the performance assessment of buffers that occur in a variety of components in telecommunication network nodes. To the best of our ability, the emphasis has been put on deriving formulae that are both efficient, accurate, and sufficiently easy to implement, rather than aspiring to justify every single step in our analyses by means of a irrefutable mathematical proof. The analyses were carried out by means of a pgf-approach, which allows us to express the quantities of interest, as far as possible, as functions of the system parameters.

We took off in Chapter II by investigating a relatively basic queuing model, that laid the foundations for some of the computational algorithms and solution techniques that proved to be useful in the remainder of the work. In this model, packet arrivals were generated by N (identical) sources, and the aggregate packet arrival process was described by means of a series of i.i.d. drvs. The packet size was set equal to 1 slot, with multiple servers (i.e., $c>1$) that provide their transmission. At first we considered an infinite-capacity buffer, and for this model, we investigated the buffer content and packet sojourn time, which ultimately led to closed-form expressions for the mean and variance of these quantities. We also provided accurate lower and upper bounds (at least if c is not too high) and a close approximation, for the terms that contain the boundary probabilities in our formulae, that reduce all numerical calculations to an absolute minimum. In addition, the asymptotic tail behaviour of the buffer content and packet sojourn time pmf and ccdf was calculated as well, based on a single dominant-pole approximation technique. Moreover, we also examined the case of a finite-capacity buffer, and presented a detailed analysis for the loss process. Specifically, we were able to express the main performance measures related this loss process in a finite-capacity buffer, in terms of the performance indices in an infinite-capacity buffer, which resulted in efficient computational procedures for calculating these quantities.

Next, in Chapter III, we then extended these results to a general framework where the N sources are all described by a D-BMAP with L (≥ 2) states. Focussing on the model with infinite storage capacity, we again derived semi-analytic closed-form expressions for the mean, variance, and tail distribution of the buffer content and the packet sojourn time. This was done by adopting a somewhat particular solution technique, based on a comprehensive use of pgfs to represent the quantities of interest and the equations that relate them. This forced us to devote some special attention to the potential – albeit removable – singularities of these pgfs that were inherent to this approach, but had the considerable advantage that results, such as the aforementioned means and variances, could be expressed, to the largest possible extent, in terms of the parameters that determine the packet arrival process. We thus found that the burst factor, represented by κ_b, plays a primary role in the evaluation of these performance indices. Moreover, the solution technique that we applied led, in an almost natural way, to a multiple-pole approximation procedure for the pmf and ccdf of the buffer content and packet sojourn time as well. Furthermore, due to the potentially huge size of the state space of the Markov chain that describes the arrival process, finding accurate approximations for the terms in our formulae that contain the boundary probabilities became even more urgent, and we indicated how this can be done in an efficient manner, that is reasonably accurate in the single-server case; for $c>1$ however (and especially for c much larger than 1), there is still some room for improvement. Finally, we extended the main results to the case of a heterogeneous D-BMAP, by focussing on the AO and FOC ordering schemes respectively. It should be noted that deriving numerical results in the AO case can become computationally demanding compared to the FOC case. This chapter was then concluded with a case study for a scenario whereby a buffer in a network node is fed by packetised telephone traffic, with the purpose of acquiring a better understanding of such a system, as well as illustrating the efficacy and usefulness of the results that were derived thus far.

Finally, in Chapter IV, we extended the analysis and results of the foregoing chapter to the case where the packet sizes were represented by a generally distributed drv. We therefore had to confine ourselves to the single-server case $c=1$, since such models with multiple servers are notoriously hard to solve. The analysis of the buffer content and packet sojourn time, yet again based on an expedient and adequate application of pgfs and their properties, was presented. Making an extensive use of the analyses techniques that were adopted thus far, we once more obtained (semi-)analytic expressions for the mean, variance, and tail distribution, of the buffer content and packet sojourn times, in the case of AO ordering, as well as for the FOC ordering algorithm. Similarly as before, numerical results are much harder to come by in the former case compared to the latter one.

In the course of this monograph, we have illustrated the main results by means of a considerable number of numerical examples that (hopefully) provided some insight in the underlying mechanisms that govern this kind of system's buffer behaviour and performance. We also encountered a number of issues that are still unresolved, or only partly resolved. Just to mention the most important ones :

- the results presented in Chapter II suggested that there is a close resemblance between the 10^{-X} quantile of a drv such as the queue content or packet waiting time on the one hand, and its moments – and the standard deviation in particular – on the other hand. This phenomenon raises a numer of questions that deserve further investigation;
- when studying a packet stream that traverses multiple ATM switching elements, generating end-to-end results, as well as results for the shared buffer memory, is still quite a challenge;
- due to the huge size of the state space (i.e., 10^6 or more in some examples) that we occasionally encounter in case of a D-BMAP, it becomes obvious that establishing close approximations for the terms that contain the boundary probabilities in our formulae for the mean, variance and tail distribution of the buffer content and packet sojourn time, is crucial. Especially in the multi-server case, a refinement of the results that have been obtained thus far would be welcomed;
- phenomena, such as the asymptotic behaviour of the buffer content and packet sojourn time quantiles for increasing values of κ_b, or the (asymptotic) behaviour in our case study of the packet waiting time quantiles as the link rate becomes higher and higher, are quite intriguing, and warrant further attention;
- in case of an i.i.d. packet arrival process, we were able to establish close and useful relationships between the quantities of interest in the case of a buffer with finite storage capacity on the one hand, and an infinite-sized buffer on the other hand. For a buffer fed by a D-BMAP however, this issue has not been fully resolved up to now;
- finding analytic solutions for multi-server models with generally distributed packet transmission times, may be one of the main challenges in this research area.

Appendix A

Discrete random variables and their pgfs

Formally, a *stochastic* or *random variable* X is defined as a measurable real-valued function from the set of outcomes (or sample space) of X, represented by Ω, to a measurable space. This measurable space is the space of possible values of the function, and it is usually taken to be the set of real numbers $\mathbb{R}$.

Consider the probability space $(\Omega, E, \Pr[\cdot])$ where the set of events E, which is a set of subsets of Ω, satisfies the properties of a σ-algebra. $\Pr[\cdot]$ is a suitable probability measure that assigns a probability to each element of E, and let the elements of the sample space Ω be represented by ω. Formally, a function $X : \Omega \rightarrow \mathbb{R}$ is a (real-valued) random variable, if for every subset $E_r=\{\omega : X(\omega) \leq r\}$ of Ω, where $r \in \mathbb{R}$, the property $E_r \in E$ also holds.

This rather formal definition enables the construction of the *cumulative distribution function* $F_X(x)$ of the random variable X, with $F_X(x)=\Pr[X \leq x]$.

A *discrete random variable* (drv), is a random variable for which the sample space Ω is a countable space. In all cases encountered in this work, the drv X will represent a *counting process* (the number of packet arrivals during a slot, the number of packets in a buffer at some given time instant, the number of slots between the arrival and departure slot of a packet, …) implying that $\Omega=\{0, 1, 2, 3, \ldots\} \equiv \mathbb{N}$. The *probability mass function* (pmf) $x(n)$ of X then is shorthand for

$$x(n) \equiv \Pr[X=n] = F_X(n)-F_X(n\text{-}1) \quad , 0 \leq n \quad ,$$

and due to the normalisation condition, it is required that

$$\sum_{n=0}^{\infty} x(n) = 1 \quad . \tag{A.1}$$

The *probability generating function* (pgf) that is associated with a drv X, which is a function of a complex variable $z \in \mathbb{C}$, is defined as the power series

$$X(z) = \mathbf{E}\left[z^X\right] = \sum_{n=0}^{\infty} z^n \Pr[X = n] \quad , \quad z \in \mathbb{C} \quad , \tag{A.2}$$

where the $\mathbf{E}[\cdot]$ operator represents the expected value of the argument, which is, in general, a function of one or more (discrete) random variables. Hence, the pgf $X(z)$ is the (complex) z-transform of the pmf $x(n)$, and the expression in the right-hand side is valid for all z in the complex plane for which the power series in the right-hand side converges. Mark that, due to the normalisation condition (A.1), the power series in the right-hand side of (A.2) converges for $z=1$. If we denote by $\mathcal{R}$ the radius of convergence of $X(z)$, then from this observation one can easily infer that $\mathcal{R} \geq 1$. For reasons that will be clarified below, we will solely deal with pgfs for which $\mathcal{R}>1$ in this dissertation. The situation where $\mathcal{R}=1$ demands an adjusted case-by-case approach, and falls outside the scope of this work.

In the following sections of this appendix, we will derive and discuss some properties of the drv X and its pgf. Let us conclude this part by pointing out that there is a one-to-one correspondence between the pmf of a drv and its pgf, i.e., each pgf is uniquely defined by the associated pmf, and vice versa. Throughout this work, it will be abundantly illustrated that carrying out calculations with pgfs, i.e., in the so-called transform or z-domain, often is advantageous compared to calculations in the probability domain.

This is exemplified by the extension of the foregoing notions and definitions to the multivariate case, where for two drvs X_1 and X_2, a *joint pmf* $\Pr[X_1=n, X_2=m]$ is now used to characterise these quantities and their mutual dependence. It is not difficult to show, for two *statistically independent* drvs X_1 and X_2 – meaning that $\Pr[X_1=n, X_2=m]=\Pr[X_1=n]\cdot\Pr[X_2=m]$, $\forall m,n$ – that their *joint pgf* then satisfies

$$\begin{aligned} X(z_1, z_2) \triangleq \mathbf{E}\left[z_1^{X_1} z_2^{X_2}\right] &= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} z_1^n z_2^m \Pr[X_1 = n, X_2 = m] \quad , \quad z \in \mathbb{C} \\ &\equiv \sum_{n=0}^{\infty} z_1^n \Pr[X_1 = n] \sum_{m=0}^{\infty} z_2^m \Pr[X_2 = m] = X_1(z_1)\, X_2(z_2) \quad , \end{aligned}$$

a property that will be extensively relied upon in the course of this monograph, and (partly) explains the latter remark concerning calculations in the z-domain.

A.1 Moments of X

The characteristics of a drv are often represented by its (central) moments. Let us, as is usually the case, denote by μ_X, σ_X^2 and $\mu_{3,X}$ the mean, variance and 3rd central moment respectively, of the drv X. These quantities can be calculated from its pmf as

$$\mu_X = \mathbf{E}[X] = \sum_{n=0}^{\infty} nx(n)$$

$$\sigma_X^2 = \mathbf{E}\left[(X-\mu_X)^2\right] = \mathbf{E}\left[X^2\right] - \mu_X^2$$
$$= \sum_{n=0}^{\infty} n^2 x(n) - \left(\sum_{n=0}^{\infty} nx(n)\right)^2 \qquad \text{(A.3)}$$

$$\mu_{3,X} = \mathbf{E}\left[(X-\mu_X)^3\right] = \mathbf{E}\left[X^3\right] - 3\mu_X \mathbf{E}\left[X^2\right] + 2\mu_X^3$$
$$= \sum_{n=0}^{\infty} n^3 x(n) - 3\left(\sum_{n=0}^{\infty} nx(n)\right)\left(\sum_{n=0}^{\infty} n^2 x(n)\right)^2 + 2\left(\sum_{n=0}^{\infty} nx(n)\right)^3 .$$

The circumstances under which the sums in these expressions converge will be commented on hereafter. For completeness, let us also define the *coefficient of variation* of a random variable as the ratio of its standard deviation versus its mean, i.e.,

$$C_{V,X} = \frac{\sigma_X}{\mu_X} \quad . \qquad \text{(A.4)}$$

This is a dimensionless number that allows the comparison of the variation of random variables that have significantly differing mean values.

Finally, in the multivariate case, the *correlation coefficient* of two random variables X_1 and X_2 is in general defined as

$$\rho_{X_1,X_2} \triangleq \frac{\mathbf{E}[X_1 \cdot X_2] - \mathbf{E}[X_1]\mathbf{E}[X_2]}{\sigma_{X_1}\sigma_{X_2}} \quad , \qquad \text{(A.5)}$$

and whenever this quantity is equal to 0, the two drvs are said to be *uncorrelated*.

A.2 Properties of $X(z)$

Let us derive a couple of properties of $X(z)$ that will turn out to be useful in the analyses presented throughout this document :

- The normalisation condition (A.1) implies that

$$X(1) = 1 \quad . \qquad \text{(A.6)}$$

- The pgf $X(z)$ is *bounded for all z that fall within the closed complex unit disk* : for values of z that belong to $\{z \in \mathbb{C} : |z| \leq 1\}$, $X(z)$ satisfies

$$|X(z)| = \left|\sum_{n=0}^{\infty} z^n x(n)\right| \le \sum_{n=0}^{\infty} |z|^n x(n) \le \sum_{n=0}^{\infty} x(n) \le 1 \ , \tag{A.7}$$

which proves this statement.

Also, since the power series in (A.2) converges for all z that fall inside the complex unit circle, the radius of convergence of any pgf $X(z)$, represented by $\mathcal{R}$, is at least 1, i.e., $\mathcal{R} \ge 1$. It can be shown that the following three properties hold (see [188]) :

- $X(z)$ has for its domain of convergence the open disk $\{ z \in \mathbb{C} : |z| < \mathcal{R} \}$;
- for any number r, $0 < r < \mathcal{R}$, the pgf $X(z)$ is analytic in the closed disk $\{ z \in \mathbb{C} : |z| \le r \}$;
- the pgf $X(z)$ may be differentiated as often as required within its domain of convergence $\{ z \in \mathbb{C} : |z| < \mathcal{R} \}$, and the differentiated series has the same domain of convergence as the original one.

Consequently, since $z=0$ falls within the domain of convergence of $X(z)$, we may differentiate $X(z)$ any number of times with respect to z for $z=0$, which induces the *probability generating property*

$$x(n) = \frac{1}{n!} \frac{d^n}{dz^n} X(z) \bigg|_{z=0} \ , \tag{A.8}$$

which, in principle, allows one to retrieve the pmf $x(n)$ from the pgf $X(z)$.

From the third previous property, it becomes clear that, if we preclude the case $\mathcal{R}=1$, i.e., if

$$\mathcal{R} > 1 \ , \tag{A.9}$$

then $z=1$ falls within the domain of convergence of $X(z)$, and we may differentiate $X(z)$ any number of times with respect to z for $z=1$, which induces the *moment generating property*

$$\frac{d^i}{dz^i} X(z) \Big|_{z=1} = \mathbf{E}\left[\prod_{j=0}^{i-1} (X - j) \right] \ . \tag{A.10}$$

This, for instance, implies that the first three (central) moments of the drv X can be calculated from the first three derivatives of $X(z)$ at $z=1$, as

$$\begin{aligned} \mu_X &= X'(1) \\ \sigma_X^2 &= X''(1) - X'(1)(X'(1) - 1) \\ \mu_{3,X} &= X'''(1) - 3(X'(1) - 1)X''(1) + X'(1)(X'(1) - 1)(2X'(1) - 1) \ , \end{aligned} \tag{A.11}$$

where we use primes as the standard shorthand notation to represent derivatives with respect to the argument of a single-valued function.

The condition (A.9) ensures that these derivatives exist, or, equivalently, that the sums in expressions (A.3) for the moments converge. Moreover, also note that under such circumstances, $X(z)$ is an analytic function at least for those values of z that satisfy $|z|\leq 1$. In this document, we will solely consider pgfs whose radius of convergence satisfies (A.9), implying that all moments of the corresponding drv are finite. For $\mathcal{R}=1$, (some of) the moments of X may be infinitely large, a situation that demands a specialized in-depth approach that depends on the specific form of $X(z)$, and that we do not wish to explore any further at this instance.

A.3 Tail distribution of X

In many environments where queuing analysis is employed to assess the system performance, we need information on the whole pmf $x(n)$ of a random variable X, rather than its first few moments. In particular the situation where n is 'large' is of interest, since this may relate to an event that rarely occurs, but potentially has grave consequences (such as a system disaster). The behaviour of $x(n)$ when n becomes large is often referred to as the *asymptotic behaviour* of $x(n)$, or the *tail distribution* of the drv X (e.g., [118]).

The probability generating property (A.8), in principle, allows the extraction of $x(n)$ from $X(z)$. However, even when $X(z)$ is know as a closed-form formula, the calculation of its n-th derivative with respect to z for $z=0$ – both analytically as well as numerically – in most cases quickly becomes unfeasible as n increases.

An often followed approach is to consider the right-hand side of (A.2) as the Laurent series expansion of the function $X(z)$, which converges for $|z|<\mathcal{R}$, and from which we can deduce that

$$x(n)=\frac{1}{2\pi\imath}\oint_{C_0}\frac{X(u)du}{u^{n+1}} \quad , \tag{A.12}$$

where the contour C_0 can be any circle centred on the origin $z=0$ and traversed in the positive direction, with radius r such that $0<r<\mathcal{R}$. The numerical calculation of this integral (frequently after some carefully chosen transformation of the variable u in this expression), is the approach of choice of the (inverse) *discrete Fourier transform* (DFT), and the *fast Fourier transform* (FFT) technique (see also [145], [201], [236]). However, these numerical calculation tools once again suffer from the drawback that they become increasingly unreliable as n becomes larger and larger.

Let us therefore follow a different approach, and from now on focus on the situation that is outlined in Appendix C, where $X(z)$ is such that its singularities stem from the zeroes of a denominator of the form $z^c-E(z)$. Then, as shown in Appendix C, provided that $c>E'(1)$ and the condition (C.4) is satisfied, the singularity with the smallest modulus of $X(z)$ is a simple

real pole, denoted by z_o, with $z_o>1$. Suppose that $X(z)$ has M isolated singularities, denoted by $z_{o,i}$, $1\leq i\leq M$, and we rearrange these quantities such that $z_o\equiv|z_{o,1}|\leq|z_{o,2}|\leq\ldots\leq|z_{o,M}|$. Now, suppose that z_o is the *only* singularity of $X(z)$ with modulus z_o, i.e., $|z_o|<|z_{o,2}|\leq\ldots\leq|z_{o,M}|$. This will, generally speaking, be the case if $\mathfrak{L}=1$, where $\mathfrak{L}$ denotes the *highest common factor* (hcf) of the set of integers $\{n : x(n)\neq 0\}$ (see the concluding remarks of Appendix C). If we take the contour integral over a circle C_R with radius R centred around $z=0$, $C_R=\{z\in\mathbb{C} : |z|=R\}$, then with $R\rightarrow\infty$, we can resort to the residue theorem to transform (A.12) into

$$x(n)=-\sum_{i=1}^{M}\mathrm{Res}\left[\frac{X(z)}{z^{n+1}},z_{o,i}\right]+I_n+\vartheta_n \quad ,$$

where I_n equals the contour integral at infinity

$$I_n=\lim_{R\rightarrow\infty}\oint_{C_R}\frac{X(z)}{z^{n+1}}dz \quad ,$$

and ϑ_n encompasses the terms that stem from branch cuts that we may need to take in the complex plain to close the contour at infinity. In this expression, $X(z)$ must be regarded upon as the analytic continuation of the power series (A.2) outside its domain of convergence. Assuming that $z_{o,i}$ is a pole of $X(z)$ with multiplicity m_i, we find that

$$\begin{aligned}\mathrm{Res}\left[\frac{X(z)}{z^{n+1}},z_{o,i}\right]&=-z_{o,i}^{-n}\frac{\left(-z_{o,i}\right)^{m-1}}{(m-1)!}\lim_{z\rightarrow z_{o,i}}\frac{d^{m_i-1}}{dz^{m_i-1}}\left(\left(1-\frac{z}{z_{o,i}}\right)^{m_i}X(z)\right)\\&\triangleq-\zeta_i z_{o,i}^{-n} \quad ,\end{aligned}$$

which leads to

$$x(n)=\sum_{i=1}^{M}\zeta_i z_{o,i}^{-n}+I_n+\vartheta_n \quad .$$

Under the assumptions described above, it is obvious that, provided that n is large enough, the sum in the above expression is dominated by the term associated with z_o. Although it is impossible to prove in general, this will also be true with respect to the value of ϑ_n, since all singularities of $X(z)$ that are not poles have a modulus larger than z_o as well. In addition, one can often prove, as will be illustrated later on by some examples, that $I_n=0$ for large enough values of n. Bearing in mind that z_o is a simple pole of $X(z)$, we thus obtain the following expression that captures the asymptotic behaviour of the pmf $x(n)$

$$x(n)\cong\zeta z_o^{-n} \quad , \tag{A.13a}$$

where $\zeta\equiv\zeta_1$ can be calculated as

$$\zeta = \lim_{z \to z_0} \left(\left(1 - \frac{z}{z_o} \right) X(z) \right) \qquad \text{(A.13b)}$$

SOME REMARKS

It is straightforward to extend the previous analysis to cover the case where $X(z)$ has multiple poles with modulus z_o. It suffices to add an appropriate term to (A.13a,b) for each of these additional poles, taking into account their respective multiplicity.

A class of interesting pgfs is the situation where $X(z)$ is a *meromorphic* function in the complex plane, i.e., $X(z)$ can be written as the ratio $f(z)/g(z)$, whereby both $f(z)$ and $g(z)$ are analytic in the complex plane (and of course $g(z)$ not a constant equal to zero). Then all singularities of $X(z)$ perforce are poles, which are the zeros of $g(z)$. If this is the case, the contour C_R can be closed without branch cuts, implying that $\vartheta_n=0$. This for instance occurs if $X(z)$ stands for the pgfs $Q(z)$ or $S(z)$ – of the steady-state queue or system contents respectively, that were calculated in Chapter II of this thesis – provided that the pgf $E(z)$, that describes the i.i.d. packet arrival process, is a meromorphic function in the complex plane as well. With respect to the evaluation of I_n, the following theorem comes in handy (see also [188]) :

THEOREM : *conditions for the vanishing of an integral around an arc*

Let $f(z)$ be a function that may or may not be analytic within and on C_R, and define $M(r) = \max_{z \in C_r} |f(z)|$, where C_r can be any arc on the circle in the complex plane around the origin with radius r, which does not intersect with a branch cut of $f(z)$. Then,

$$\lim_{R \to \infty} RM(R) = 0 \Rightarrow \lim_{R \to \infty} \oint_{C_R} f(z)dz = 0 \quad . \qquad \square$$

If $X(z)$ does not have branch cuts in the complex plane, then from this theorem and the definition of I_n, we can distil the following criterion for which $I_n=0$:

$$\lim_{R \to \infty} \left| X(R.e^{\imath\theta}) \right| R^{-n} = 0 \quad , \quad \forall \theta \in [0, 2\pi[\quad , \quad \Rightarrow \quad I_n = 0 \quad . \qquad \text{(A.14)}$$

Let us illustrate this for the pgf $Q(z)$, representing the steady-state pgf of the queue content, that was derived in Chapter II :

$$Q(z) = c(1-\rho) \frac{(z-1)}{z^c - E(z)} \Psi(z) \quad ,$$

with $\Psi(z)$ a polynomial of degree c-1; we also point out that the coefficient of z^{c-1} in this polynomial is equal to $\Pr[s \le c-1]/(c(1-\rho))$, where the drv s represents the system content, and where the load ρ satisfies $\rho < 1$.

If we let $E(z)$ represent a *Poisson* arrival process with arrival rate $c\rho$, i.e., $E(z) = \exp\{c\rho(z-1)\}$, then the behaviour of $\left|Q(R.e^{\imath\theta})\right|$ on C_R depends on whether or not $\cos(\theta)>0$, leading to, for $0 \le \theta < 2\pi$,

$$\lim_{R\to\infty} \left|Q(R.e^{\imath\theta})\right| = \begin{cases} 0 & , \ \theta \in [0, \pi/2[\cup]3\pi/2, 2\pi[\\ \Pr[s \le c-1] & , \ \theta \in [\pi/2, 3\pi/2] \end{cases} .$$

Consequently, whatever the value of θ, this implies that

$$\lim_{R\to\infty} \left|Q(R.e^{\imath\theta})\right| R^{-1} = 0 \quad , \quad 0 \le \theta < 2\pi \quad .$$

Therefore, in view of (A.14), we conclude that I_n=0, for all $n \ge 1$.

As a second example, consider the case that $E(z)$ is a *rational* function, with $E'(1)<c$. Let d_N and d_D be the degree of the numerator and denominator respectively, and let a_N (a_D) denote the coefficient of z^{d_N} (z^{d_D}) in the numerator (denominator). Then one can easily verify that, for $0 \le \theta < 2\pi$,

$$\lim_{R\to\infty} \left|Q(R.e^{\imath\theta})\right| = \begin{cases} 0 & , \ d_N > d_D + c \\ \dfrac{\Pr[s \le c-1]}{|1 - a_N/a_D|} & , \ d_N = d_D + c \\ \Pr[s \le c-1] & , \ d_N < d_D + c \end{cases} .$$

For completeness, note that d_N=d_D+c, a_N=a_D would cause some of the probabilities $\Pr[e=n]$ to be either negative or larger than 1. Consequently,

$$\lim_{R\to\infty} \left|Q(R.e^{\imath\theta})\right| R^{-1} = 0 \quad , \quad 0 \le \theta < 2\pi \quad ,$$

once again leading to the conclusion that I_n=0 for all $n \ge 1$.

These examples illustrate that in a lot of interesting cases, the tail approximation (A.13a,b) will provide accurate results, as soon as the values of n are such that $|z_{o,i}|^{-n}$, $2 \le i \le M$, can be neglected compared to z_o^{-n}. Throughout this work, the validity and accuracy of such an approach (and the ones derived thereof) will be confirmed by means of a considerable amount of numerical examples.

Appendix B

Rouché's theorem and its application

Rouché's theorem can be formulated as follows (e.g. [188]) :

Let $f(z)$ and $g(z)$, $z \in \mathbb{C}$, be analytic inside a simply connected bounded domain $\mathcal{D}$ in the complex plane, as well as on its boundary $\mathcal{C}$, and let these functions be such that

$$|f(z)| > |g(z)| \quad , \tag{B.1}$$

on the boundary $\mathcal{C}$.

Then the functions $f(z)$ and $f(z)+g(z)$ have the same number of zeros in $\mathcal{D}$, if each zero is counted according to its multiplicity.

We will now apply this to prove that the equation

$$z^c - E(z) = 0 \quad ,$$

has c solutions inside the closed complex unit disk $\{z \in \mathbb{C} : |z| \le 1\}$ – including $z=1$ – with $E(z)$ a pgf that describes the number of packet arrivals per slot, and by assumption satisfies

$$E'(1) < c \quad .$$

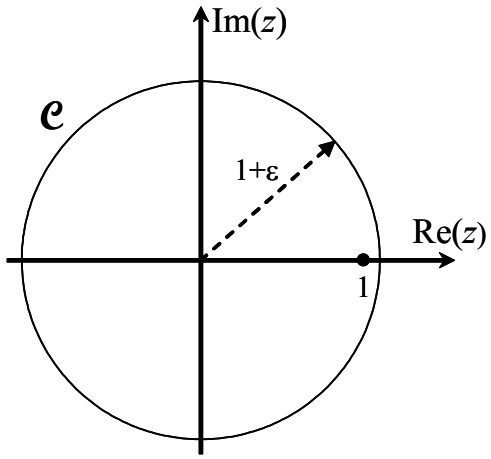


Figure B-1 : contour $\mathcal{C}$ for applying Rouché's theorem

Adopting the above notations, we set $f(z)=z^c$ and $g(z)=-E(z)$. Furthermore, as shown in Fig. B-1 consider the boundary $\mathcal{C}$ as the circle around the origin $z=0$ with radius $1+\varepsilon$, i.e.,

$$\mathcal{C}=\{z \in \mathbb{C} : |z|=1+\varepsilon\} \quad ,$$

for infinitesimal small values of the positive real number ε. Under the condition that $\mathcal{R}$, the radius of convergence of $E(z)$, satisfies (A.9), then ε can be chosen such that the domain $\mathcal{D}$ bounded by $\mathcal{C}$ falls within the domain of convergence of $E(z)$, implying that $g(z)$ is an analytic function in $\mathcal{D}$ and on $\mathcal{C}$; and obviously so is $f(z)$. We now prove inequality (B.1) :

$$|1+\varepsilon|^c \overset{?}{>} |E(1+\varepsilon)| < \sum_{i=0}^{\infty} |1+\varepsilon|^i \Pr[e=i] \quad ,$$

which, for sufficiently small values of ε, will be satisfied if

$$1+c\varepsilon \overset{?}{>} \sum_{i=0}^{\infty} (1+i\varepsilon)\Pr[e=i] = 1+\varepsilon E'(1) \quad ,$$

and in view of $E'(1) < c$, this inequality is indeed valid.

We can therefore conclude that the function $z^c - E(z)$ has c solutions within $\mathcal{C}$. Apparently, $z=1$ is one of these solutions. In general, this will be the only solution with modulus equal to 1 (see also the closing remark of Appendix C). Letting $\varepsilon \to 0$, this implies that $z^c - E(z)$ has c-1 zeros in the open complex unit disk $\{z \in \mathbb{C} : |z|<1\}$.

Appendix C

In search of the dominant pole

Let us consider the equation

$$z^c - E(z) = 0 \quad , \quad z \in \mathbb{C} \tag{C.1}$$

where $E(z)$ represents the pgf of an i.i.d. packet arrival process, i.e.,

$$E(z) = \sum_{i=0}^{\infty} z^i \Pr[e = i] \quad . \tag{C.2}$$

We consider pgfs $E(z)$ for which the corresponding i.i.d. packet arrival process from time to time generates more than c arrivals, i.e.,

$$\{n \in \mathbb{N} : n > c \wedge \Pr[e = n] \neq 0\} \neq \phi \quad . \tag{C.3}$$

Observe that his condition in fact expresses that the queuing system under consideration in Chapter II is a non-trivial one.

We also assume that the equilibrium condition (I.4), implying that $E'(1) = \mu_e < c \Leftrightarrow \rho < 1$, is satisfied. Let $\mathcal{R}$ be the radius of convergence of $E(z)$, with $\mathcal{R} > 1$ by assumption (see (A.9) in Appendix A and related comments), and suppose that the following inequality is valid :

$$\lim_{\substack{z \to \mathcal{R} \\ <}} E(z) \big/ z^c > 1 \quad . \tag{C.4}$$

Note that if this condition is satisfied, then (C.3) automatically holds as well.

To set our mind, we first take a look at the situation where $\mathcal{R}$ is finite. We define the real-valued function $F(z)$, $z \in \mathbb{R}$, on the open interval $]1, \mathcal{R}[$, as

$$F(z) = \frac{z^c - E(z)}{z^c (z-1)} \ , \ z \in]1, \mathcal{R}[\quad . \tag{C.5}$$

Since $E(z)$ is a continuous function in the open interval $]1,\mathcal{R}[$ (which immediately follows from the series expansion (C.2), and the fact that $]1,\mathcal{R}[$ falls within its domain of convergence), so is $F(z)$. Evidently, the roots of $F(z)$ in $]1,\mathcal{R}[$ are also zeroes of (C.1) in this region, and vice versa.

We can now prove the following properties :

PROPOSITION 1 : $F(z)$ has at least one zero $z_o > 1$ in the interval $]1,\mathcal{R}[$

Since $F(z)$ is continuous within $]1,\mathcal{R}[$, and

$$F(1) = c(1-\rho) > 0$$
$$\lim_{\substack{z\to\mathcal{R} \\ <}} F(z) < 0 \quad ,$$

where the latter inequality follows from (C.4), we can invoke *Bolzano's theorem*, which states that, under these circumstances, there is at least one $u \in]1,\mathcal{R}[$ that satisfies $F(u)=0$. □

Let us denote by z_o any of these solutions.

PROPOSITION 2 : z_o is the only zero of $F(z)$ in the interval $]1,\mathcal{R}[$

This immediately follows from the observation that $F(z)$ is a strictly decreasing function for $z \in]1,\mathcal{R}[$. Indeed, due to expressions (C.2) and (C.5), $F(z)$ can be rewritten as

$$F(z) = -\sum_{i=0}^{\infty} \frac{z^{i-c}-1}{z-1}\Pr[e=i] = \sum_{i=0}^{c-1} z^{-1}\frac{z^{-(c-i)}-1}{z^{-1}-1}\Pr[e=i] - \sum_{i=c}^{\infty}\frac{z^{i-c}-1}{z-1}\Pr[e=i]$$
$$= \sum_{i=0}^{c-1}\sum_{j=1}^{c-i} z^{-j}\Pr[e=i] - \sum_{i=c}^{\infty}\sum_{j=0}^{i-c-1} z^{j}\Pr[e=i] \ .$$

Because values $z \in]1,\mathcal{R}[$ fall within the domain of convergence of $F(z)$, these sums (and their derivatives) converge for these particular values of z, and we may conclude that

$$F'(z) = -\left(\sum_{i=0}^{c-1}\sum_{j=1}^{c-i} jz^{-j-1}\Pr[e=i] + \sum_{i=c}^{\infty}\sum_{j=0}^{i-c-1} jz^{j-1}\Pr[e=i]\right) < 0 \ , \ z \in]1,\mathcal{R}[\ .$$ □

Note that at least one of the terms in the sums between parentheses necessarily is nonzero, in view of (C.3).

PROPOSITION 3 : z_o is a zero of $F(z)$ with multiplicity 1

From $F(z_o)$, expression (C.5) for $F(z)$, and $F'(z_o)$, we deduce that

$$\begin{aligned} cz_o^{c-1} - E'(z_o) &= z_o^c(z_o - 1)F'(z_o) \\ &< 0 \ , \end{aligned} \tag{C.6}$$

which proves that z_o is a zero of $z^c - E(z)$ with multiplicity 1; otherwise the equality $cz_o^{c-1} - E'(z_o) = 0$ would hold. □

PROPOSITION 4 : Equation (C.1) has no solutions outside the complex unit disk with modulus less than z_o

In order to prove this property, we appeal to Rouché's theorem discussed in Appendix B, with $f(z)=z^c$ and $g(z)=-E(z)$.

As depicted in Fig. C-1, let us consider the boundary $\mathcal{C}$ as the circle around the origin $z=0$ with radius z_o-ε, i.e.,

$$\mathcal{C}=\{z\in\mathbb{C} \ : |z|=z_o-\varepsilon\} \quad ,$$

for infinitesimal small values of the positive real number ε. Since $z_o>1$, we are sure that we can choose ε such that z_o-ε>1. Note that, as the domain $\mathcal{D}$, bounded by $\mathcal{C}$, falls within the domain of convergence of $E(z)$, $g(z)$ is an analytic function in $\mathcal{D}$ and on $\mathcal{C}$, and obviously so is $f(z)$. What remains to be proven is inequality (B.1) :

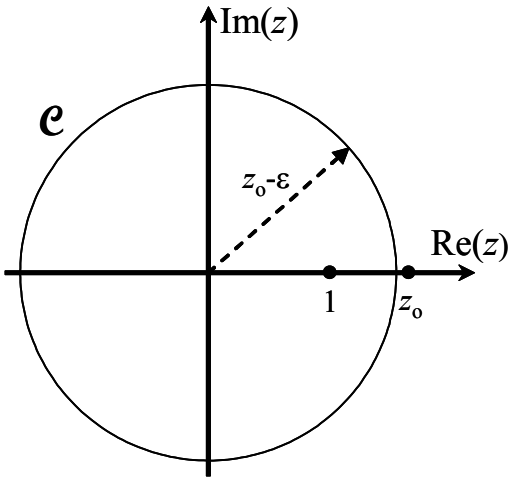


Figure C-1: contour $\mathcal{C}$ for applying Rouché's theorem

$$|z_o - \varepsilon|^c \overset{?}{>} |E(z_o - \varepsilon)| < \sum_{i=0}^{\infty} |z_o - \varepsilon|^i \Pr[e = i] \quad ,$$

which, for sufficiently small values of ε, will be satisfied if

$$z_o^c - c\varepsilon z_o^{c-1} \overset{?}{>} \sum_{i=0}^{\infty} \left(z_o^i - i\varepsilon z_o^{i-1} \right) \Pr[e = i] = E(z_o) - \varepsilon E'(z_o) \quad .$$

As z_o is a zero of $z^c - E(z)$, this reduces to the condition

$$cz_o^{c-1} < E'(z_o) \quad ,$$

which we know to be valid, e.g. equation (C.6). □

This proposition is confirmed by *Vivanti's theorem* :

> If τ is the radius of convergence of a power series $X(z)$ with real and positive coefficients, then $z=\tau$ necessarily is a singularity of $X(z)$.

Indeed, if the equation (C.1) would have a solution outside the complex unit disk with modulus less than z_o, this would be a singularity for the pgf (say $Q(z)$) of which $z^c - E(z)$ is the denominator. Hence, the radius of convergence of $Q(z)$ would be less than z_o, and from Vivanti's theorem we deduce that there would be another solution of (C.2) in the interval $]1,z_o[$. We have proved by means of Proposition 2 that the latter is false; hence, equation (C.2) has no solution outside the complex unit disk with modulus less than z_o.

CONCLUSIONS

Summarizing, taking into account the abovementioned properties, we conclude that z_o is the only solution of (C.1) that is located outside the complex unit disk and falls within the region of convergence of $E(z)$. The root z_o is a positive real number, and has multiplicity 1.

In this document, we will only deal with pgfs $E(z)$ that satisfy the inequality (C.4), implying that these results are of interest for any pgf (say $Q(z)$) for which z^c-$E(z)$ is the denominator. Provided that all singularities of $Q(z)$ stem from the zeros of z^c-$E(z)$, this means that the zero $z=z_o$ will be the dominant singularity of $Q(z)$. Its radius of convergence is equal to z_o as well, and its domain of convergence coincides with the open disk with radius z_o, i.e., $\{z \in \mathbb{C} : |z| < z_o\}$.

SOME REMARKS

- Up to now, we have tacitly assumed that $\mathcal{R}$ is finite. Nevertheless, it is not difficult to check that the above derivations and properties do not drastically change when the radius of convergence of $E(z)$ becomes infinite. The zero z_o then is the only solution of (C.1) on the positive real axis with modulus larger than 1.
- The above derivations are valid under quite general conditions :
  - $E(z)$ has a radius of convergence $\mathcal{R}>1$, i.e., we exclude the case where $\mathcal{R}=1$;
  - the inequality (C.4) holds.
- Let us briefly take a look at some situations where (C.4) is not valid. This will either be the case when $E(z)$ is a polynomial of degree no larger than c – a situation that we preclude provided that (C.4) (and hence (C.3)) is satisfied – or whenever $z=\mathcal{R}$ is a branch point of $E(z)$ with $\mathcal{R}^c \leq E(\mathcal{R})$. Such pgfs exist; a class that for instance could fit this description is given by

$$E(z) = 1 + ac\rho(\mathcal{R} - 1)\left[1 - \left(\frac{\mathcal{R} - z}{\mathcal{R} - 1}\right)^{1/a}\right] \quad , \quad a>1 \quad , \tag{C.7a}$$

where the parameters (c, ρ, a, $\mathcal{R}$) must satisfy

$$1 + ac\rho(\mathcal{R} - 1)\left[1 - \left(\frac{\mathcal{R}}{\mathcal{R} - 1}\right)^{1/a}\right] > 0 \quad , \tag{C.7b}$$

in order to ensure that all probabilities that correspond to the pgf in (C.7) are positive quantities. This situation is illustrated in Fig. C-2, where we have plotted $F(z)/c$ versus z for c=1,2,3,4, with $E(z)$ in the expression for $F(z)$ given by (C.7a), for parameter values (ρ, a, $\mathcal{R}$)=(0.4, 5, 1.5). If c=1,2, the condition (C.4) is satisfied and $F(z)/c$ has a zero in the interval $]1, \mathcal{R}[$, as we observe from the intersect of these two curves with the abscis. For c=3,4, the condition (C.4) is no longer satisfied, and the function $F(z)/c$ does not have a zero within $]1, \mathcal{R}[$; consequently, neither will $z^c - E(z)$. In this case, the branch point $z=\mathcal{R}$ will be the dominant singularity of any pgf for which $z^c - E(z)$ is the denominator, and determining the tail behaviour of the associated pmf requires a specific case-by-case approach.

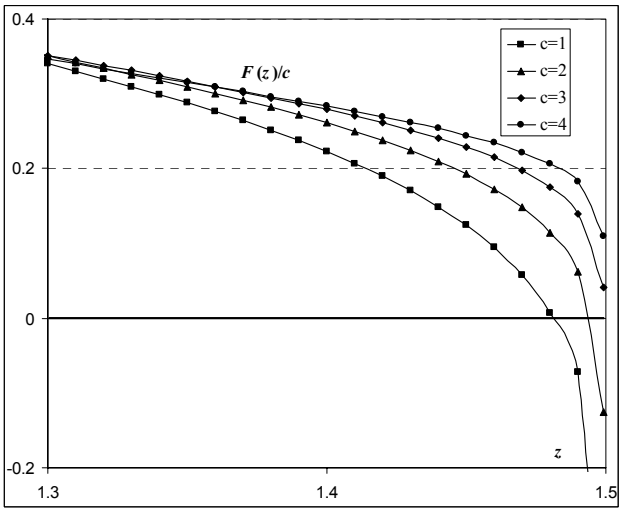


Figure C-2 : $F(z)/c$ versus z

- It is, in theory, possible that the equation (C.1) has solutions different from z_o but with modulus equal to z_o. This can for instance be the case if $\mathcal{L}$>1, where we define $\mathcal{L}$ to be the hcf of the set of integers $\left\{\{c\} \cup \left\{n \in \mathbb{N} : \Pr[e = n] \neq 0\right\}\right\}$. Indeed, since $z^c - E(z)$ can be regarded as a function of $z^{\mathcal{L}}$ under these circumstances, then if z_o is a zero of $z^c - E(z)$, so will $z_{o,k} = z_o.\exp\{2\pi\imath k/\mathcal{L}\}$, $1 \le k \le \mathcal{L}$-1, and these zeros obviously have a modulus equal to z_o. Generally speaking though, if $\mathcal{L}$=1, then z_o will be the only zero with modulus z_o.

Appendix D

The eigenvalues and –vectors of $\mathbf{Q}(z)$

D.1 calculation and general properties

As became clear from the analysis throughout Chapters III and IV, an important role is played by the eigenvalues and –vectors of the pgm $\mathbf{Q}(z)$. Let us denote by $\mathbf{\Lambda}(z)$ the $L\times L$ diagonal matrix containing the eigenvalues $\lambda_i(z)$ of a general $\mathbf{Q}(z)$, given by (III.1), which are the solutions of the *characteristic equation*

$$|\mathbf{Q}(z)-\lambda(z)\mathbf{I}|=0 \Leftrightarrow \begin{vmatrix} p_{11}G_{11}(z)-\lambda(z) & p_{12}G_{12}(z) & \cdots & p_{1L}G_{1L}(z) \\ p_{21}G_{21}(z) & p_{22}G_{22}(z)-\lambda(z) & \cdots & p_{2L}G_{2L}(z) \\ \vdots & \vdots & \ddots & \vdots \\ p_{L1}G_{L1}(z) & p_{L2}G_{L2}(z) & \cdots & p_{LL}G_{LL}(z)-\lambda(z) \end{vmatrix}=0 \quad ,$$

where $\mathbf{I}$ represents the $L\times L$ diagonal identity matrix. This characteristic equation can be formally written as

$$\sum_{l=0}^{L} C_{L-l}(z)\lambda(z)^{l}=0 \quad , \tag{D.1}$$

where, without loss of generality, we can set $C_0(z)\equiv 1$. The functions $C_l(z)$, $1\le l\le L$, are linear combinations of the $l\times l$ minors of $\mathbf{Q}(z)$, with $C_L(z)\equiv\pm|\mathbf{Q}(z)|$. Let us denote by $\mathcal{R}_{lm}$ (with $\mathcal{R}_{lm}>1$ by assumption) the radius of convergence of $G_{lm}(z)$ and define

$$\mathcal{R}\triangleq\min_{l,m}\{\mathcal{R}_{lm}\} \quad .$$

We will refer to $\mathcal{R}$ as the radius of convergence of the pgm $\mathbf{Q}(z)$, in the sense that each of the entries of $\mathbf{Q}(z)$ is an analytic function in the open disk $\{z\in\mathbb{C}:|z|<\mathcal{R}\}$. Consequently, the $C_l(z)$'s will be analytic functions in the region $\{z\in\mathbb{C}:|z|<\mathcal{R}\}$ as well. For these values of z, the characteristic equation can be regarded upon as a polynomial of degree L (with finite coefficients) for the unknown variable $\lambda(z)$, and from the fundamental theorem of algebra, we

know that this equation has exactly L solutions. Consequently, $\mathbf{\Lambda}(z)$ can be uniquely determined for all $\{z\in\mathbb{C}:|z|<\mathcal{R}\}$, and since the $C_l(z)$'s have no poles within this area, neither will $\lambda_l(z)$, $1\leq l\leq L$. However, this does not prohibit the $\lambda_l(z)$ of having *branch points* for certain values of z within this region, a matter that will be commented upon in Section D.2. Since, for $z=1$, $\mathbf{Q}(1)$ is the stochastic matrix of the modulating Markov chain that is recurrent and aperiodic by assumption (and therefore ergodic), this implies that one of the corresponding eigenvalues, say $\lambda_1(z)$, satisfies $\lambda_1(1)=1$, while the others fulfil the inequality (see [164])

$$|\lambda_l(1)|<1\ ;\ 2\leq l\leq L\ . \tag{D.2}$$

In the remainder, $\lambda_1(z)$ will be referred to as the *Perron-Frobenius* (PF)-eigenvalue of $\mathbf{Q}(z)$.

In addition, let us also define $\mathbf{W}(z)$ as the $L\times L$ matrix containing the respective $1\times L$ left-eigenvectors $\bar{\boldsymbol{w}}_l(z)$ with eigenvalue $\lambda_l(z)$ of $\mathbf{Q}(z)$, and $\mathbf{U}(z)$ as the inverse matrix of $\mathbf{W}(z)$, meaning that these matrices satisfy the following equations :

$$\begin{cases}\mathbf{W}(z)\mathbf{Q}(z)=\mathbf{\Lambda}(z)\mathbf{W}(z)\\ \mathbf{U}(z)\mathbf{W}(z)=\mathbf{W}(z)\mathbf{U}(z)=\mathbf{I}\end{cases}\ . \tag{D.3a}$$

(with $\mathbf{I}$ the $L\times L$ diagonal matrix with all elements equal to 1) implying that

$$\mathbf{Q}(z)\mathbf{U}(z)=\mathbf{U}(z)\mathbf{\Lambda}(z)\quad , \tag{D.3b}$$

which shows that $\mathbf{U}(z)$ is the $L\times L$ matrix whose columns $\bar{\boldsymbol{u}}_i(z)$ are the $L\times 1$ right-eigenvectors, with eigenvalue $\lambda_i(z)$, of $\mathbf{Q}(z)$. The respective left- and right-eigenvectors can be determined from these relations upon some constant factor, which can be fixed by requiring that the rows of either $\mathbf{W}(z)$ or $\mathbf{U}(z)$ are normalised (the former implies the latter, and vice versa)

$$\mathbf{W}(z)\bar{\boldsymbol{I}}=\bar{\boldsymbol{I}}\ \Leftrightarrow\ \mathbf{U}(z)\bar{\boldsymbol{I}}=\bar{\boldsymbol{I}}\quad , \tag{D.3c}$$

with $\bar{\boldsymbol{I}}$ the $L\times 1$ column vector with all entries equal to 1. We then also obtain

$$\mathbf{Q}(z)=\mathbf{U}(z)\mathbf{\Lambda}(z)\mathbf{W}(z)\quad . \tag{D.3d}$$

We need to emphasise that, although the solution for the eigenvectors that follows from the set of equations (D.3a-c) works well for arbitrary values of z, it nevertheless degenerates if $z=1$. First notice that the equations for $\bar{\boldsymbol{w}}_1(1)$, namely $\bar{\boldsymbol{w}}_1(1)\mathbf{Q}(1)=\bar{\boldsymbol{w}}_1(1)$ and $\bar{\boldsymbol{w}}_1(1)\bar{\boldsymbol{I}}=1$ are exactly the ones that determine the stationary vector $\bar{\boldsymbol{\pi}}^T$ defined in Section III.2.1. Also, we easily deduce that $\bar{\boldsymbol{u}}_1(1)=\bar{\boldsymbol{I}}$ is the right eigenvector of $\mathbf{Q}(1)$, if we require that $\bar{\boldsymbol{w}}_1(1)\bar{\boldsymbol{u}}_1(1)=1$

(see also [164]). However, for $2 \leq l \leq L$, and $\bar{\boldsymbol{w}}_l(1)$ being the left eigenvector associated with $\lambda_l(1)$ of the stochastic matrix $\mathbf{Q}(1)$ (with $\mathbf{Q}(1)\bar{\boldsymbol{I}} = \bar{\boldsymbol{I}}$), we can establish the relation

$$\bar{\boldsymbol{w}}_l(1)\mathbf{Q}(1) = \lambda_l(1)\bar{\boldsymbol{w}}_l(1) \;\Rightarrow\; \bar{\boldsymbol{w}}_l(1)\bar{\boldsymbol{I}} = \lambda_l(1)\bar{\boldsymbol{w}}_l(1)\bar{\boldsymbol{I}} \quad .$$

Since, for $2 \leq l \leq L$, the non-PF eigenvalue $\lambda_l(1)$ satisfies the inequality (D.2), the above relation can only hold if $\bar{\boldsymbol{w}}_l(1)\bar{\boldsymbol{I}} = 0$, which contradicts the equality $\bar{\boldsymbol{w}}_l(1)\bar{\boldsymbol{I}} = 1$ that follows from the normalisation requirement (D.3c). This explains why the entries of the solution for $\bar{\boldsymbol{w}}_l(z)$ that derives from (D.3a,c) tend to $\pm\infty$ for $z \to 1$. As a consequence, the relations (D.3a) then also indicate that the entries of $\bar{\boldsymbol{u}}_l(z)$ tend to 0, for $2 \leq l \leq L$. Summarising, we thus obtain

$$\begin{cases} \bar{\boldsymbol{u}}_1(1) = \bar{\boldsymbol{I}} \\ \bar{\boldsymbol{u}}_l(1) = \bar{\boldsymbol{0}} \;, 2 \leq l \leq L \end{cases} ; \quad \begin{cases} \bar{\boldsymbol{w}}_1(1) = \bar{\boldsymbol{\pi}}^T \\ \bar{\boldsymbol{w}}_l(1) = \bar{\boldsymbol{\infty}}^T \;, 2 \leq l \leq L \end{cases} , \tag{D.4}$$

where $\bar{\boldsymbol{0}}$ and $\bar{\boldsymbol{\infty}}$ are $L \times 1$ column vectors whose L entries are equal to 0 and $\pm\infty$ respectively. Although this solution for $2 \leq l \leq L$ seems impossible to work with at first sight, we will not shun away from it. The reason is that, throughout the derivations presented in this work, we do not make use of the value of the individual components of $\bar{\boldsymbol{w}}_l(z)$ (and $\bar{\boldsymbol{u}}_l(z)$), but rather of specific combinations of these quantities, such as the ones in equations (D.3a-d). For instance, as a small example, although the elements on the l-th row of $\mathbf{W}(z)$ tend to infinity for $2 \leq l \leq L$ and $z \to 1$, the sum of each row still equals 1 under these circumstances, i.e., $\bar{\boldsymbol{w}}_l(1)\bar{\boldsymbol{I}} = 1$, $2 \leq l \leq L$. Also, as indicated in Section III.2.1, if we denote by $u_{ij}(z)$ the entries of $\mathbf{U}(z)$ (and hence, the i-th component of $\bar{\boldsymbol{u}}_j(z)$), then observe that the identities

$$\begin{cases} \left(u_{ij}(z) \cdot \bar{\boldsymbol{w}}_j(z)\right)\mathbf{Q}(z) = \lambda_j(z)\left(u_{ij}(z) \cdot \bar{\boldsymbol{w}}_j(z)\right) \\ \left(u_{ij}(z) \cdot \bar{\boldsymbol{w}}_j(z)\right)\bar{\boldsymbol{I}} = u_{ij}(z) \end{cases} , \tag{III.7e}$$

show that $u_{ij}(z) \cdot \bar{\boldsymbol{w}}_j(z)$ is a left-eigenvector of $\mathbf{Q}(z)$ with eigenvalue $\lambda_j(z)$, and the solution of this set of equations indeed does yield a finite value at $z = 1$. Our calculations in Chapter III (and IV) will reveal that the products $u_{ij}(z) \cdot \bar{\boldsymbol{w}}_j(z)$ are exactly the ones we need in our derivations concerning the buffer analysis. In Section D.4 of this Appendix, we will highlight the specific case $L = 2$, from which it will become clear that the eigenvalues and –vectors still satisfy each of the relations of (D.3a-d) for $z \to 1$. Our calculations will further reveal that the solution for the eigenvectors that is proposed here leads to some considerable simplifications in the representation of the results that are generated throughout this monograph, which explains our preference for this somewhat particular approach.

D.2 Branch points of the eigenvalues

In view of the characteristic equation (D.1), the first derivative of the eigenvalues of $\mathbf{Q}(z)$ can be calculated in an unambiguous way from

$$\lambda_i'(z) = \frac{\sum_{l=1}^{L} C'_{L-l}(z)\lambda_i(z)^l}{\sum_{l=1}^{L} lC_{L-l}(z)\lambda_i(z)^{l-1}} \; ; \; 1 \le i \le L \quad . \tag{D.5}$$

Since the $C_l(z)$ functions are analytic for all $\{z \in \mathbb{C} : |z| < \mathcal{R}\}$, then the $\lambda_i(z)$, viewed as a function of z, are differentiable, and hence analytic in this area, except for *the zeroes of the denominator* in this expression. These values of z are the branch points of the corresponding eigenvalues; let z^* be such a value, and assume that $z^* \in \{z \in \mathbb{C} : |z| < \mathcal{R}\}$. Notice that the requirement of the denominator in the above expression being zero, is equivalent to the property that $\lambda_i(z^*)$ is a solution of the characteristic equation with multiplicity larger than 1, i.e., at least one value of j, $1 \le j \le L$, can be found for which

$$\lambda_i(z^*) = \lambda_j(z^*) \; ; \; j \ne i \quad .$$

These remarks reconfirm the observations that were also made in [163], [198], namely that the $\lambda_i(z)$'s, $\{z \in \mathbb{C} : |z| < \mathcal{R}\}$, are analytic functions in this region, *except for those values of z for which they are not distinct*.

Such (either real or complex) values z^* exist, and may even fall within the complex unit circle. To illustrate this, it suffices to focus on a small example, whereby $\mathbf{Q}(z)$ is a 2×2 matrix of the following form :

$$\mathbf{Q}(z) = \begin{bmatrix} p_1 G(z) & (1-p_1) \\ (1-p_2)G(z) & p_2 \end{bmatrix} \quad ,$$

with $0 < p_1, p_2 < 1$, and $G(z)$ any pgf that describes the number of packets that are generated when a source visits state S_1. Observe that if $p_1 + p_2 \ne 1$, this is not an i.i.d. process, i.e., there is dependence between the number of packets that are generated during subsequent slots. From the results that are presented in Section D.4 for a general two-state D-BMAP arrival process, we can deduce that $\lambda_1(z^*) = \lambda_2(z^*)$ implies that

$$p_1^2 G(z)^2 - 2(p_1 p_2 - 2(1-p_1)(1-p_2))G(z) + p_2^2 = 0 \quad .$$

As an example, for $p_1 = p_2 = 0.375$, this yields the two solutions $G(z^*) = -9$ and $G(z^*) = -1/9$ respectively. If $G(z)$ for instance represents a Bernoulli arrival process with rate γ, then one

can readily check that for $\gamma>5/9$, one of these solutions for the branch points z^* falls within the complex unit circle. The observation that these branch points may be situated inside the complex unit disk already points to the fact that these are not essential singularities of pgfs such as $Q(z)$ and $S(z)$ that describe the steady-state queue and system content, since these are perforce analytic functions for $|z|<1$. This issue will be discussed next in more detail.

To make matters even worse, inspection of results such as the ones presented in Section D.4 for L=2 reveals that these branch points of the eigenvalues are also singularities (i.e., poles) of the corresponding right-eigenvectors. Nevertheless, we can prove that these branch points are *removable singular points* (see [188]) of the main results that were derived throughout Chapter III and IV, such as the steady-state pgfs $Q(z)$ and $S(z)$. Considering L=3, we will demonstrate this for $Q(z)$, satisfying (III.25), by showing that both $Q(z^*)$ and $Q'(z^*)$ exist and can be calculated in an unambiguous way. The arguments that are hereafter developed for L=3, are exemplary for any arbitrary value of L.

To set our mind, consider a point in the complex plane that falls within the region of convergence of $\mathbf{Q}(z)$, $z^* \in \{z \in \mathbb{C} : |z| < \mathcal{R}\}$. Hence, a (finite) solution for $\lambda_i(z^*)$, $1 \leq i \leq 3$, exists, and assume that

$$\lambda_1(z^*) \neq \lambda_2(z^*) = \lambda_3(z^*) \quad ,$$

i.e., z^* is a branch point of $\lambda_2(z)$ and $\lambda_3(z)$, but not of $\lambda_1(z)$. Hence, this implies that $\lambda_1(z)$ is differentiable at $z=z^*$, and from relation (D.3b), we deduce that the same holds for the corresponding right-eigenvalue $\overline{\boldsymbol{u}}_1(z)$. The value $z=z^*$ is a pole of $\overline{\boldsymbol{u}}_2(z)$ and $\overline{\boldsymbol{u}}_3(z)$, but from the normalisation requirement $\mathbf{U}(z)\overline{\boldsymbol{1}} = \overline{\boldsymbol{1}}$, we can still deduce that the following limits

$$s_{0,i}(z^*) \triangleq \lim_{z \to z^*} \left(u_{i2}(z) + u_{i3}(z)\right) = 1 - u_{i1}(z^*)$$

$$s_{1,i}(z^*) \triangleq \lim_{z \to z^*} \frac{d}{dz}\left(u_{i2}(z) + u_{i3}(z)\right) = -u'_{i1}(z^*) \quad ,$$

exist. First, taking into account that $\lambda_2(z^*)=\lambda_3(z^*)$, then the steady-state pgf $Q(z)$ of the queue content can, for $z=z^*$, be written as (with $\overline{\boldsymbol{m}} = [m_1 \;\; m_2 \;\; m_3]^T$, and $m_3 = N - m_1 - m_2$)

$$Q(z^*) = \sum_{m_1=0}^{N} \frac{(z^*-1)\lambda_1(z^*)^{m_1}\lambda_2(z^*)^{N-m_1}}{z^{*c} - \lambda_1(z^*)^{m_1}\lambda_2(z^*)^{N-m_1}} \lim_{z \to z^*} \sum_{m_2=0}^{N-m_1} \Psi_{\overline{\boldsymbol{m}}}(z) \quad ,$$

and in view of expression (III.18a) for $\Psi_{\overline{\boldsymbol{m}}}(z)$, this quantity can be uniquely determined if

$$f_{0,\overline{\boldsymbol{l}}}(m_1) \triangleq \lim_{z \to z^*} \sum_{m_2=0}^{N-m_1} F_{\overline{\boldsymbol{l}}\overline{\boldsymbol{m}}}(z) \quad , \; \forall\, \overline{\boldsymbol{l}} \quad ,$$

exists. From definition (III.17a) for $F_{\bar{l}\bar{m}}(z)$, we can conclude that $f_{0,\bar{l}}(m_1)$ (with $\bar{l}=[l_1\ l_2\ l_3]^T$ and $l_3=N-l_1-l_2$) is the coefficient of $x_1^{m_1}$ in the expression

$$\lim_{z\to z^*}\prod_{i=1}^{3}\left(u_{i1}(z)x_1+[u_{i2}(z)+u_{i3}(z)]\right)^{l_i}=\prod_{i=1}^{3}\left(u_{i1}(z^*)x_1+s_{0,i}(z^*)\right)^{l_i} \quad ,$$

which, in view of the previous considerations, yields a finite and unambiguous result.

In an analogous way, one can show that $Q'(z)$ exists for $z=z^*$, if the following limits exist :

$$f_{1,\bar{l}}(m_1)\triangleq\lim_{z\to z^*}\frac{d}{dz}\sum_{m_2=0}^{N-m_1}F_{\bar{l}\bar{m}}(z)$$

$$\varepsilon_{\bar{l}}(m_1)\triangleq\lim_{z\to z^*}\sum_{m_2=0}^{N-m_1}\frac{d}{dy}\left(\lambda_2(y)^{m_2}\lambda_3(y)^{N-m_1-m_2}\right)\Big|_{y=z}F_{\bar{l}\bar{m}}(z) \quad , \qquad \forall\,\bar{l} \quad .$$

Based on the definitions (and existence) of $s_{0,i}(z^*)$ and $s_{1,i}(z^*)$, the existence of $f_{1,\bar{l}}(m_1)$ can be demonstrated by an identical derivation as above, so let's focus on the calculation of $\varepsilon_{\bar{l}}(m_1)$. Again, due to the definition of $F_{\bar{l}\bar{m}}(z)$, this quantity is the coefficient of $x_1^{m_1}$ in the formula

$$\lim_{z\to z^*}\frac{d}{dy}\left(\prod_{i=1}^{3}\left(u_{i1}(z)x_1+[u_{i2}(z)\lambda_2(y)+u_{i3}(z)\lambda_3(y)]\right)^{l_i}\right)\Big|_{y=z} \quad .$$

In view of the identity $\mathbf{U}(z)\mathbf{\Lambda}(z)\bar{I}=\mathbf{Q}(z)\bar{I}$ (which also holds for $z\to z^*$), we may write

$$\lim_{z\to z^*}[u_{i2}(z)\lambda_2(z)+u_{i3}(z)\lambda_3(z)]=\sum_{j=1}^{3}\left(p_{ij}G_{ij}(z^*)\right)-u_{i1}(z^*)\lambda_1(z^*)$$

$$\lim_{z\to z^*}\frac{d}{dy}([u_{i2}(z)\lambda_2(y)+u_{i3}(z)\lambda_3(y)])\Big|_{y=z}=\sum_{j=1}^{3}\left(p_{ij}G'_{ij}(z^*)\right)+\lambda_1(z^*)s_{1,i}(z^*)-u_{i1}(z^*)\lambda'_1(z^*) \quad ,$$

and since the expressions in the right-hand side of these two formulae can be unambiguously calculated and are finite for $z^*\in\{z\in\mathbb{C}:|z|<\mathcal{R}\}$, the limits in the left-hand side exist, and hence, so does the limit in the penultimate expression.

Summarising, $Q(z)$ and its first derivative exist for $z=z^*$, and the same approach can be adopted to show that the n-th order derivative of $Q(z)$ exists. Consequently, although $z=z^*$ is a branch point of two of the eigenvalues, and a pole of the corresponding right-eigenvectors, it is not a singularity of $Q(z)$, but a so-called *removable singular point*. Clearly, the same can be said of all branch points of the eigenvalues that fall within the region $\{z\in\mathbb{C}:|z|<\mathcal{R}\}$. Similar arguments can be developed for other pgfs of interest (such as the steady-state pgf $S(z)$ of the system content), and/or other values of L.

D.3 Derivatives of the PF-eigenvalues and –vectors

In the derivations concerning the mean and variance of such drvs as the queue and system content, it became obvious that the derivatives with respect to z for $z=1$ of the PF-eigenvalues (and right-eigenvectors) play a primary part in the results that we eventually obtain. In this section, we briefly outline a procedure that allows us to calculate these quantities.

We will commence by showing that $\lambda_1'(1)$ is equal to the mean number of packet arrivals during an arbitrary slot (generated by a single source). We start by proving that

LEMMA :

$$\overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{u}}_i'(1)=0 \; ; \; 1\le i\le L \quad . \tag{D.6}$$

First, due to $\mathbf{U}(z)\overline{\boldsymbol{1}}=\overline{\boldsymbol{1}}$, we have that $\mathbf{U}'(z)\overline{\boldsymbol{1}}=\overline{\boldsymbol{0}}$. Consequently, we may write

$$\sum_{i=1}^{L} \overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{u}}_i'(1)=0 \quad . \tag{*}$$

On the other hand, taking into account that $\overline{\boldsymbol{u}}_i(1)=\overline{\boldsymbol{0}}$ for $2\le i\le L$, we deduce from the eigenvalue equation $\mathbf{Q}(z)\overline{\boldsymbol{u}}_i(z)=\lambda_i(z)\overline{\boldsymbol{u}}_i(z)$ that

$$\overline{\boldsymbol{\pi}}^T \mathbf{Q}(1)\overline{\boldsymbol{u}}_i'(1)=\lambda_i(1)\overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{u}}_i(1) \Leftrightarrow (1-\lambda_i(1))\overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{u}}_i'(1)=0 \; ; \; 2\le i\le L \quad .$$

In view of the property (D.2), this leads to the conclusion that

$$\overline{\boldsymbol{\pi}}^T \overline{\boldsymbol{u}}_i'(1)=0 \; ; \; 2\le i\le L \quad .$$

Making use of equation (*), this implies that this equality also holds for i=1; consequently, (D.6) is indeed valid. □

Then, from the matrix relation $\mathbf{Q}(z)\mathbf{U}(z)=\mathbf{U}(z)\mathbf{\Lambda}(z)$, we derive that

$$\begin{aligned} &\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)\mathbf{U}(1)\overline{\boldsymbol{1}}+\overline{\boldsymbol{\pi}}^T \mathbf{Q}(1)\mathbf{U}'(1)\overline{\boldsymbol{1}}=\overline{\boldsymbol{\pi}}^T \mathbf{U}'(1)\mathbf{\Lambda}(z)\overline{\boldsymbol{1}}+\overline{\boldsymbol{\pi}}^T \mathbf{U}(1)\mathbf{\Lambda}'(z)\overline{\boldsymbol{1}} \\ \Leftrightarrow \quad &\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)\overline{\boldsymbol{1}}+\overline{\boldsymbol{\pi}}^T \mathbf{U}'(1)\overline{\boldsymbol{1}}=\overline{\boldsymbol{\pi}}^T \mathbf{U}'(1)\mathbf{\Lambda}(z)\overline{\boldsymbol{1}}+\lambda_1'(1) \quad , \end{aligned}$$

which, in view of the previous lemma indeed reduces to

$$\lambda_1'(1)=\overline{\boldsymbol{\pi}}^T \mathbf{Q}'(1)\overline{\boldsymbol{1}}=c\rho/N \quad , \tag{D.7a}$$

in accordance to expression (III.11b). Note that this result can, in principle, also be obtained from (D.5) with $\lambda_1(1)=1$.

Higher-order derivatives of the PF-eigenvalue can then be computed by taking

successive derivatives of the characteristic equation (D.1). We, for instance, obtain for the second-order derivative

$$\lambda_1''(1) = -\frac{\sum_{l=1}^{L}\left(C''_{L-l}(1) + 2\frac{c\rho}{N} l C'_{L-l}(1) + \left(\frac{c\rho}{N}\right)^2 l(l-1) C_{L-l}(1)\right)}{\sum_{l=1}^{L} l C_{L-l}(1)} \quad , \tag{D.7b}$$

and a similar procedure can be followed for the calculation of $\lambda_1'''(1)$ that is required in the expressions for the variance of the queue and system content. Also remember that we have already established that the value $z=1$ is not a branch point of $\lambda_1(z)$, which implies that the denominator in the above expression is perforce nonzero.

The first- and second-order derivatives of the PF-eigenvector with respect to z for $z=1$ can then be calculated from taking the first two derivatives at $z=1$ of the characteristic equation $\mathbf{Q}(z)\overline{\boldsymbol{u}}_1(z) = \lambda_1(z)\overline{\boldsymbol{u}}_1(z)$, which leads to

$$\begin{aligned} (\mathbf{Q}(1)-\mathbf{I})\overline{\boldsymbol{u}}_1'(1) &= \lambda_1'(1)\overline{\boldsymbol{I}} - \mathbf{Q}'(1)\overline{\boldsymbol{I}} \\ (\mathbf{Q}(1)-\mathbf{I})\overline{\boldsymbol{u}}_1''(1) &= \lambda_1''(1)\overline{\boldsymbol{I}} + 2(\lambda_1'(1)\mathbf{I} - \mathbf{Q}'(1))\overline{\boldsymbol{u}}_1'(1) - \mathbf{Q}''(1)\overline{\boldsymbol{I}} \end{aligned} \quad ,$$

i.e., the calculation of $\overline{\boldsymbol{u}}_1'(1)$ and $\overline{\boldsymbol{u}}_1''(1)$ each time requires solving a set of L linear equations for the L unknowns.

The expressions for the variance of the queue and system content also contain the first-order derivatives of the non-PF-eigenvalues and –vectors with respect to z for $z=1$. These can be computed in a completely analogous way as above for the PF-eigenvalue and –vector.

D.4 The two-state D-BMAP arrival process

Taking into account expression (III.1) for $\mathbf{Q}(z)$ (with $[q_{ij}(z)] \equiv [p_{ij}G_{ij}(z)]$), the characteristic equation of this $L \times L$ matrix in case of $L=2$ can be written as

$$\lambda(z)^2 - (q_{11}(z) + q_{22}(z))\lambda(z) + q_{11}(z)q_{22}(z) - q_{12}(z)q_{21}(z) = 0 \quad .$$

Solution of this characteristic equation leads to

$$\left.\begin{matrix}\lambda_1(z) \\ \lambda_2(z)\end{matrix}\right\} = \frac{q_{11}(z) + q_{22}(z) \pm \left((q_{11}(z) - q_{22}(z))^2 + 4q_{12}(z)q_{21}(z)\right)^{1/2}}{2} \quad , \tag{D.8a}$$

where the $\lambda_1(z)$ corresponds to the '+'-sign in the numerator. Note that

$$\lambda_1(1) = 1$$
$$\lambda_2(1) = 1 - p_{12} - p_{21} \quad ,$$

which indicates that $\lambda_1(z)$ is the PF-eigenvalue of **Q**(z), introduced in the preceding sections.

In addition, solving (D.3a,b) yields the following expression for **W**(z) :

$$\mathbf{W}(z) = \begin{bmatrix} \dfrac{q_{21}(z)}{\lambda_1(z) - q_{11}(z) + q_{21}(z)} & \dfrac{\lambda_1(z) - q_{11}(z)}{\lambda_1(z) - q_{11}(z) + q_{21}(z)} \\ \dfrac{q_{21}(z)}{\lambda_2(z) - q_{11}(z) + q_{21}(z)} & \dfrac{\lambda_2(z) - q_{11}(z)}{\lambda_2(z) - q_{11}(z) + q_{21}(z)} \end{bmatrix} \quad . \tag{D.8b}$$

Inversion of this matrix yields expressions for the components of **U**(z) :

$$\mathbf{U}(z) = \begin{bmatrix} \dfrac{\big(\lambda_1(z) - q_{11}(z) + q_{21}(z)\big) q_{12}(z)}{\big(\lambda_1(z) - q_{11}(z)\big)\big(\lambda_1(z) - \lambda_2(z)\big)} & \dfrac{\big(\lambda_2(z) - q_{11}(z) + q_{21}(z)\big) q_{12}(z)}{\big(\lambda_2(z) - q_{11}(z)\big)\big(\lambda_2(z) - \lambda_1(z)\big)} \\ \dfrac{\big(\lambda_1(z) - q_{11}(z) + q_{21}(z)\big)}{\big(\lambda_1(z) - \lambda_2(z)\big)} & \dfrac{\big(\lambda_2(z) - q_{11}(z) + q_{21}(z)\big)}{\big(\lambda_2(z) - \lambda_1(z)\big)} \end{bmatrix} \quad . \tag{D.8c}$$

One can now check that these results for **U**(z) and **W**(z) fulfil each of the expressions (D.3a-d), in particular for z=1, which is a pole of the components of $\bar{\boldsymbol{w}}_2(z)$, and values z=z* that satisfy $\lambda_1(z^*)=\lambda_2(z^*)$, which are branch points of these eigenvalues and poles of the right eigenvectors of **Q**(z). In addition, note that the components of $u_{ij}(z) \cdot \bar{\boldsymbol{w}}_j(z)$ are indeed finite, as stated in the comments concerning (III.7e).

In our expressions for the mean and variance of drvs such as the queue and system content, we require the first, second and third order derivatives with respect to z for z=1 of the PF-eigenvalue. Taking the appropriate derivatives of the characteristic equation with $\lambda_1(1)$=1, we obtain (since all derivatives are taken with respect to z for z=1, we temporarily omit the explicit reference to the argument of the derivatives of these functions)

$$\lambda_1' = \sum_{i=1}^{2} \pi_i \sum_{j=1}^{2} q_{ij}' = c\rho/N \quad , \tag{D.9a}$$

in accordance to (D.7a). Furthermore, we find

$$\begin{aligned} \lambda_1'' &= \sum_{i=1}^{2} \pi_i \sum_{j=1}^{2} q_{ij}'' + 2\frac{\lambda_1'\lambda_2' + q_{12}'q_{21}' - q_{11}'q_{22}'}{p_{12} + p_{21}} \\ \lambda_1''' &= \sum_{i=1}^{2} \pi_i \sum_{j=1}^{2} q_{ij}''' + 3\frac{\lambda_1''\lambda_2' + \lambda_1'\lambda_2'' + q_{12}''q_{21}' + q_{12}'q_{21}'' - q_{11}''q_{22}' - q_{11}'q_{22}''}{p_{12} + p_{21}} \end{aligned} \tag{D.9b}$$

The n-th derivative of $\lambda_2(z)$ with respect to z for $z=1$ on the other hand, is readily calculated in terms of the n-th derivative of $\lambda_1(z)$, from the relation $\lambda_1(z)+\lambda_2(z)=q_{11}(z)+q_{22}(z)$.

Similarly, we also require the first and second-order derivatives with respect to z for $z=1$ of the PF right-eigenvector. First of all, we point out that in view of (D.8c), $\mathbf{U}(1)$ satisfies

$$\mathbf{U}(1)=\begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix} \quad ,$$

in concurrence with (D.4). Then, taking the appropriate derivatives of the equations that follow from (D.3b,c) and the relation $\mathbf{Q}(z)\mathbf{U}(z)=\mathbf{U}(z)\mathbf{\Lambda}(z)$, we obtain

$$\mathbf{U}'=\frac{\beta_0}{p_{12}}\begin{bmatrix}-\pi_2 & \pi_2\\ \pi_1 & -\pi_1\end{bmatrix}$$

$$\mathbf{U}''=\frac{\beta_1}{p_{12}}\begin{bmatrix}-\pi_2 & \pi_2\\ \pi_1 & -\pi_1\end{bmatrix}+\frac{\beta_2}{p_{12}}\begin{bmatrix}-\pi_2 & \pi_2\\ -\pi_2 & \pi_2\end{bmatrix} \quad , \tag{D.9c}$$

where

$$\begin{aligned}\beta_0&=\lambda_1'-(q_{11}'+q_{12}')\\ \beta_1&=\lambda_1''-2\frac{\beta_0 p}{p_{12}+p_{21}}-(q_{11}''+q_{12}'')-2(q_{11}'u_{11}'+q_{12}'u_{21}')\\ \beta_2&=-2\frac{\beta_0\lambda_2'}{p_{12}+p_{21}}+2(q_{11}'u_{11}'+q_{12}'u_{21}') \quad .\end{aligned} \tag{D.9d}$$

Bibliography

THE AUTHOR'S LIST OF PUBLICATIONS

Journal papers

[1] H. Bruneel, B. Steyaert, 'Buffer requirements for ATM switches with multiserver output queues', *Electronics Letters*, 1991, vol. 27(8), pp. 671-672.

[2] H. Bruneel, B. Steyaert, E. Desmet, G.H. Petit, 'An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queues', *International Journal of Digital and Analog Communication Systems*, 1992, vol. 5, pp. 193-201.

[3] B. Steyaert, H. Bruneel, G.H. Petit, E. Desmet, 'End-to-end delays in multistage ATM switching networks : approximate analytic derivation of tail probabilities', *Computer Networks and ISDN Systems*, vol. 25(11), 1993, pp. 1227-1241.

[4] B. Steyaert, H. Bruneel, 'Exact message delay derivation for TDMA schemes with multiple contiguous outputs and general independent arrivals', *International Journal of Electronics and Communication* (AEU), vol. 47(2), 1993, pp. 77-84.

[5] Y. Xiong, B. Steyaert, H. Bruneel, 'An ATM statistical multiplexer with on/off sources and spacing : numerical and analytical performance studies', *Performance Evaluation*, vol. 21(1), 1994, pp. 37- 58.

[6] H. Bruneel, B. Steyaert, E. Desmet, G.H. Petit, 'Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues', *European Journal of Operational Research*, vol. 76, 1994, pp. 563-572.

[7] B. Steyaert, H. Bruneel, 'On the performance of multiplexers with three-state bursty sources : analytical results', *IEEE Transactions on Communications*, vol. 43(2-4), 1995, pp. 1299-1303.

[8] B. Steyaert, H. Bruneel, 'Queueing delay of an individual cell stream in an ATM multiplexer : a generating-functions approach', *Journal on Communications*, vol. XLVII(1), 1996, pp. 57-62.

[9] Y. Xiong, H. Bruneel, B. Steyaert, 'Deriving delay characteristics from queue length statistics in discrete-time queues with multiple servers', *Performance Evaluation*, vol. 24(3), 1996, pp. 189-204.

[10] H. Bruneel, B. Steyaert, 'Storage requirements in ATM switching elements with correlated arrivals and independent uniform routing', *Performance Evaluation*, vol. 25(3), 1996, pp. 193-209.

[11] B. Steyaert, H. Bruneel, 'Delay performance of CBR traffic in an ATM multiplexer with high-priority background traffic', *International Journal of Computer Systems Science & Engineering*, vol. 11(6), 1996, pp. 393-399.

[12] B. Steyaert, Y. Xiong, 'Analysis of a discrete-time queue with general three-state Markovian traffic sources', *Performance Evaluation*, vol. 24(4), 1996, pp. 277-294.

[13] H. Bruneel, V. Inghelbrecht, B. Steyaert, 'Buffering for a transmission rate reduction by a rational factor', *Electronics Letters*, vol. 33(7), 1997, pp. 550-551.

[14] B. Steyaert, Y. Xiong, H. Bruneel, 'An efficient solution technique for discrete-time queues fed by heterogeneous traffic', *International Journal of Communication Systems*, vol. 10, 1997, pp. 73-86.

[15] P. Van Mieghem, B. Steyaert, G.H. Petit, 'Performance of cell loss priority management schemes in a single server queue', *International Journal of Communication Systems*, vol. 10, 1997, pp. 161-180.

[16] K. Kang, B. Steyaert, C. Kim, 'A simple relation between loss performance and buffer contents in a statistical multiplexer with periodic vacations', *IEICE Transactions on Communications*, vol. E80-B(11), 1997, pp. 1749-1752.

[17] K. Kang, B. Steyaert, 'Bound analysis for WRR scheduling in a statistical multiplexer with bursty sources', *Telecommunication Systems*, vol. 12(2-3), 1999, pp. 123-147.

[18] D. De Vleeschauwer, G. Petit, S. Wittevrongel, B. Steyaert, H. Bruneel, 'An accurate closed-form formula to calculate the dejittering delay in packetised voice transport', *Lecture Notes in Computer Science*, vol. 1815, 2000, pp. 374-385.

[19] J. Walraevens, B. Steyaert, H. Bruneel, 'Analysis of packet delay in a *GI*-*G*-1 queue with non-preemptive priority scheduling', *Lecture Notes in Computer Science*, vol. 1815, 2000, pp. 433-445.

[20] V. Inghelbrecht, B. Steyaert, H. Bruneel, S. Wittevrongel, 'Rate adapters with bursty arrivals and rational rate reduction : queueing analysis', *Performance Evaluation*, vol. 42(1), 2000, pp. 41-56.

[21] J. Walraevens, B. Steyaert, H. Bruneel, 'Performance analysis of the system contents in a discrete-time non-preemptive priority queue with general service times', *Journal of the Belgian Operations Research Society* (JORBEL), vol. 40(1-2), 2000, pp. 91-103.

[22] V. Inghelbrecht, B. Steyaert, H. Bruneel, S. Wittevrongel, 'Buffer contents and cell delay in a rate adaptation buffer with Markovian arrivals', *Computers & Operations Research*, vol. 28(9), 2001, pp. 885-898.

[23] D. Fiems, B. Steyaert, H. Bruneel, 'Performance evaluation of CAI and RAI transmission modes in a GI-G-1 queue', *Computers & Operations Research*, vol. 28(13), 2001, pp. 1299-1313.

[24] D. De Vleeschauwer, G.H. Petit, B. Steyaert, S. Wittevrongel, H. Bruneel, 'Calculation of end-to-end delay quantile in network of *M*/*G*/1 queues', *Electronics Letters*, vol. 37(8), 2001, pp.535-536.

[25] D. Fiems, and B. Steyaert, H. Bruneel, 'Randomly interrupted *GI*-*G*-1 queues : service strategies and stability issues', *Annals of Operations Research*, vol. 112, 2002, pp. 171-183.

[26] J. Walraevens, B. Steyaert, H. Bruneel, Performance analysis of a *GI*-*G*-1 preemptive resume priority buffer, *Lecture Notes in Computer Science*, vol. 2345, 2002, pp. 745-756.

[27] J. Walraevens, B. Steyaert, H. Bruneel, 'Delay characteristics in discrete-time *GI*-*G*-1 queues with non-preemptive priority queueing discipline', *Performance Evaluation*, vol. 50(1), 2002, pp. 53-75.

[28] J. Walraevens, B. Steyaert, H. Bruneel, 'Performance analysis of a single-server ATM queue with a priority scheduling', *Computers & Operations Research,* vol. 30, 2003, pp. 1807-1829.

[29] D. Fiems, B. Steyaert, H. Bruneel, 'Analysis of a discrete-time GI-G-1 queueing model subjected to bursty interruptions', *Computers & Operations Research*, vol. 30, 2003, pp. 139-153.

[30] J. Walraevens, B. Steyaert, H. Bruneel, 'Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline', *European Journal of Operational Research*, vol. 157(1), 2004, pp. 130-151.

[31] G. Van Hoey, D. De Vleeschauwer, B. Steyaert, V. Inghelbrecht, H. Bruneel, 'Benefit of admission control in aggregation network dimensioning for video services', *Lecture Notes in Computer Science*, vol. 3042, 2004, pp. 357-368.

[32] D. Fiems, B. Steyaert, H. Bruneel, 'Discrete-time queues with generally distributed service times and renewal-type server interruptions', *Performance Evaluation*, vol. 55(3-4), 2004, pp. 277-298.

[33] B. De Schepper, B. Steyaert, S. Wittevrongel, H. Bruneel, 'Influence of the Time slot Interchange Mechanism on the buffer behavior of an integrated switching element', *IEICE Transactions on Communications*, 2004, vol. E87-B(4), pp. 909-917.

[34] V. Inghelbrecht, B. Steyaert, S. Wittevrongel, H. Bruneel, 'An analytical approach to obtain the interdeparture time characteristics in a multistage VoIP network', *Telecommunication Systems*, vol. 27(1), 2004, pp. 33-45.

[35] J. Walraevens, B. Steyaert, H. Bruneel, 'A packet switch with a priority scheduling discipline: performance analysis', *Telecommunication Systems*, vol. 28(1), 2005, pp. 53-77.

[36] V. Inghelbrecht, B. Steyaert, S. Wittevrongel, H. Bruneel, 'Burst loss and delay in optical buffers with offset-time management', *Telecommunication Systems*, vol. 31(2-3), 2006, pp. 247-258.

[37] J. Walraevens, B. Steyaert, M. Moeneclaey, H. Bruneel, 'A discrete-time HOL priority queue with multiple traffic classes', *Lecture Notes in Computer Science*, vol. 3420, 2005, pp. 620-627.

[38] J. Walraevens, B. Steyaert, H. Bruneel, 'A preemptive repeat priority queue with resampling: performance analysis', *Annals of Operations Research*, vol. 146, 2006, pp. 189-202.

[39] J. Walraevens, B. Steyaert, M. Moeneclaey, H. Bruneel, 'Delay analysis of a HOL priority queue', *Telecommunication Systems*, vol. 30(1-3), 2005, pp. 81-98.

[40] B. De Schepper, B. Steyaert, S. Wittevrongel, M. Moeneclaey, H. Bruneel, 'Constant hardware delay in integrated switching elements with multiserver output queues', *IEE Proc. Communications*, vol. 153(5), 2006, pp. 664-670.

[41] J. Walraevens, B. Steyaert, H. Bruneel, 'Analysis of a discrete-time preemptive resume priority buffer', *European Journal of Operational Research*, 2007.

[42] B. Steyaert, K. Laevens, D. De Vleeschauwer, H. Bruneel, 'Analysis and design of a playout buffer for VBR streaming video', *Annals of Operations Research*, 2007.

Contribution to books

[43] B. Steyaert, H. Bruneel, 'Accurate approximation of the cell loss ratio in ATM buffers with multiple servers', *Performance Modelling and Evaluation of ATM Networks, Volume 1*, Chapman & Hall, London, 1995 (ISBN : 0-412-71140-0), pp. 285-296.

[44] V. Inghelbrecht, B. Steyaert, H. Bruneel, S. Wittevrongel, 'Rate adapters with bursty arrivals and rational rate reduction : queuing analysis', *System Performance Evaluation, Methodologies and Applications,* CRC Press, Boca Raton, U.S.A., 2000 (ISBN : 0-8493-2357-6), pp. 3-22.

[45] J. Walraevens, B. Steyaert, H. Bruneel, 'Delay analysis of a discrete-time non-preemptive priority buffer with 3 traffic classes', *Recent Advances in Communications and Computer Science, WSEAS Press*, Athens (ISBN: 960-8052-86-6), 2003, pp. 350-357.

Conference papers and contributions

[46] B. Steyaert, H. Bruneel, 'A general analysis of the packet delay in TDMA channels with contiguous slot assignments', Proc. *International Conference on Communications* (ICC '91; Denver, Colo., June 1991), pp. 1539-1543.

[47] E. Desmet B. Steyaert, H. Bruneel, G.H. Petit, 'Tail distributions of queue length and delay in discrete-time multiserver queueing models applicable in ATM networks', *Queueing Performance and Control in ATM*, Proc. *13th International Teletraffic Congress* (ITC-13; Copenhagen, June 1991), pp. 1-6.

[48] B. Steyaert, H. Bruneel, 'Approximate calculation of overflow probabilities in a shared-buffer packet switch', Abstract Booklet *11th European Congress on Operational Research*, (EURO XI; Aachen, July 16-19, 1991), pp. 229-230.

[49] B. Steyaert, H. Bruneel, 'An effective algorithm to calculate the distribution of the buffer contents and the packet delay in a multiplexer with bursty sources', Proc. *IEEE Global Telecommunications Conference* (IEEE GLOBECOM '91; Phoenix, Az., Nov. 1991), pp. 471-475.

[50] B. Steyaert, H. Bruneel, 'Analytical study of the buffer contents in an ATM switch with Markovian arrivals', Proc. *International Seminar on Teletraffic and Networks* (ISTN-92; Beijing, Sept. 1992), pp. 9-12.

[51] B. Steyaert, H. Bruneel, 'Analysis of ATM switching modules with channel grouping in a bursty-source environment', Proc. Confe*rence on Integrated Broadband Communication Networks and Services* (IBCN&S; Copenhagen, Apr. 1993), pp. 24.1-1 - 24.1-12.

[52] B. Steyaert, H. Bruneel, 'Analysis of the buffer behavior of an ATM switch with cell arrivals generated by a general correlated process: a novel approach', Proc. *St.-Petersburg Regional International Teletraffic Seminar on Digital Communication Network Management* (St.-Petersburg, June 1993), pp. 199-217.

[53] Y. Xiong, B. Steyaert, 'Performance analysis of an ATM switching node with three-state Markovian traffic generators and random spacing, High Speed Networks and their Performance', Proc. *5th IFIP WG6.4 International Conference on Data Communication Systems and Performance* (Raleigh, NC; Oct. 1993), pp.63-82.

[54] B. Steyaert, H. Bruneel, Y. Xiong, 'A general relationship between buffer occupancy and delay in discrete-time multiserver queueing models applicable in ATM networks', Proc. *IEEE Conference on Computer Communications* (IEEE INFOCOM '93; San Francisco, March-Apr. 1993), pp. 1250-1258.

[55] B. Steyaert and H. Bruneel, 'Buffer contents and delay in a statistical multiplexer with Markovian arrivals', Proc. *2nd International Conference on Computer Communications and Networks* (IC^{3}N; San Diego, Ca., June 1993), pp. 426-430.

[56] B. Steyaert, H. Bruneel, 'Analytic derivation of the cell loss probability in finite multiserver buffers from infinite-buffer results', Proc. *2nd IFIP WG6.4 Workshop on Performance Modelling and Evaluation of ATM Networks* (Bradford, UK; July 1994), pp. 18/1-18/11.

[57] B. Steyaert, H. Bruneel, 'Delay characteristics of a tagged cell stream in a discrete-time single-server queue in the presence of background traffic', Proc. *IEEE Global Telecommunications Conference* (IEEE GLOBECOM '94; San Francisco, Ca., Nov. 1994), pp. 1118-1122.

[58] E. Desmet, B. Steyaert, H. Bruneel, G.H. Petit, 'Performance analysis of a resequencing unit in a multipath self-routing switch fabric', *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Proc. *14th International Teletraffic Conference*, (ITC-14; Antibes, Juan-les-Pins, June 1994), Volume 1a, pp. 611-621.

[59] B. Steyaert, H. Bruneel, 'Analytic performance evaluation of ATM queues with Markov modulated arrivals and multiple servers', Proc. *COST 242 Management Committee Meeting* (Cambridge, U. K., 18-19 January 1995).

[60] B. Steyaert, H. Bruneel, H. Michiel, G.H. Petit, 'Impact of the input-output read-write cycle on the buffer characteristics of an ATM switching element with shared memory', Proc. *First IEEE International Workshop on Broadband Switching Systems* (IEEE BSS '95; Poznan, Apr. 19-21, 1995), pp. 140-154.

[61] B. Steyaert, Y. Xiong, 'Buffer requirements in ATM-related queueing models with bursty traffic : an alternative approach', *Data Communications and their Performance*, Proc. *6th IFIP WG6.3 Conference on Performance of Computer Networks* (Istanbul, Turkey, Oct. 1995), pp. 165-178.

[62] B. Steyaert, H. Bruneel, 'Tagged cell stream delay in a single-server queue with delay-priority background traffic', Proc. *International Teletraffic Seminar* (Bangkok, Nov. 28 - Dec. 1, 1995), pp. 46.0-46.11.

[63] B. Steyaert, H. Bruneel, 'Buffer dimensioning for rate adaptation modules, Proc. of the International IFIP-IEEE Conference on Broadband Communications', Proc. *Broadband Communications '96* (Montreal, 23-25 April 1996), pp. 489-500.

[64] B. Steyaert, H. Bruneel, 'Moments and tail distribution of queue lengths and cell delays in multiserver ATM buffers' Presentation *Second IFIP WG 6.3 Workshop on Performance of Communication Systems* (Balatonfured, Oct. 12-15, 1996).

[65] B. De Schepper, B. Steyaert, H. Bruneel, 'Cell resequencing in an ATM switch, Proceedings COST 257 Management Committee Meeting (Espoo, Finland, 27-28 May 1997).

[66] B. Steyaert, H. Bruneel, 'Cell delay and queue length characteristics in ATM buffers with multiple output lines', Proc. *IEEE ATM'97 Workshop* (Lisbon, Portugal, 25-28 May 1997), pp. 611-620.

[67] P. Van Mieghem, G.H. Petit, B. Steyaert, 'Throughput optimality of single queue priority schemes', Proc. *2nd IEEE Symposium on Computers and Communications* (Alexandria, Egypt, July 1997), pp. 240-248.

[68] V. Inghelbrecht, H. Bruneel, B. Steyaert, 'Queueing analysis of a rate adaptation buffer', Proc. *Fifth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks* (Ilkley, July 21-23, 1997), Part 2 : Research Papers, pp. 32/1-32/8.

[69] B. De Schepper, B. Steyaert, H. Bruneel, 'Restoring cell integrity in an ATM switch by means of delay equalisation : queueing analysis', Proc. *Fifth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks* Ilkley, July 21-23, 1997), Part 2 : Research Papers, pp. 37/1-37/10.

[70] B. Steyaert, H. Bruneel, 'Analysis of a rate adaptation buffer for small rate mismatches' Presentation *Third IFIP WG 6.3 Workshop on Performance of Communication Systems* (Gent, Aug. 20-22, 1997).

[71] K. Kang, B. Steyaert, C. Kim, 'An approximation for the buffer behavior of a statistical multiplexer with contiguous-slot assignments and deterministic vacations', Proc. *IFIP 13th International Conference on Computer Communication* (ICCC'97; Cannes, Nov. 1997), pp. 407-412.

[72] B. Steyaert, K. Kang, H. Bruneel, 'Performance of a Round-Robin bandwidth allocation scheme', Proc. *COST 257 Management Committee Meeting* (Rome, Italy, Jan. 15-16 1998).

[73] V. Inghelbrecht, B. Steyaert, H. Bruneel, S. Wittevrongel, 'Buffer behavior of a rate adapter with correlated arrivals', Proc. *Sixth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks* (Ilkley, July 20-22, 1998), Part 2 : Research Papers, pp. 64/1-64/10.

[74] V. Inghelbrecht, B. Steyaert, H. Bruneel, S. Wittevrongel, 'Transmission rate reduction of correlated traffic by a rational factor', Proc. *COST 257 Management Committee Meeting* (Warsaw, Poland, May 17-18, 1999).

[75] K. Kang, B. Steyaert, C. Kim, 'Service separation in ATM networks using a hardware efficient rate-controlled cell multiplexer', Proc. *16th International Teletraffic Congress* (ITC-16; Edinburgh, June 7-11, 1999), pp. 395-404.

[76] V. Inghelbrecht, B. Steyaert, H. Bruneel, 'Characteristics of switched traffic in a multistage ATM network', Proc. *Seventh IFIP Workshop on Performance Modelling and Evaluation of ATM Networks* (Antwerpen, June 28-30, 1999), pp. 1/13-13/13.

[77] B. Steyaert, H. Bruneel, 'Finite versus infinite capacity multiserver ATM buffers with correlated arrivals : analytic derivation of the CLR' Presentation *Fourth IFIP WG 6.3 Workshop on Performance of Communication Systems* (Rethymnon, Crete, Aug 29 – Sept 1, 1999).

[78] D. Fiems, B. Steyaert, H. Bruneel, 'Buffer contents and message delay for a *GI-G*-1 queueing model with random server vacations' Proc. *COST 257 Management Committee Meeting* (Larnaca, Cyprus, Sep. 30 – Oct. 1, 1999).

[79] V. Inghelbrecht, B. Steyaert, H. Bruneel, S. Wittevrongel, 'Rate adapters with bursty arrivals and rational rate reduction : queueing analysis' Presentation *IFIP WG 7.3 Conference on Modeling and Performance Evaluation of Computer Systems and Networks*, (PERFORMANCE '99; Istanbul, Oct. 15-17, 1999).

[80] J. Walraevens, B. Steyaert, H. Bruneel, 'Analysis of the system contents in a *GI-G*-1 queue with non-preemptive priority scheduling', Book of Abstracts *Fourteenth Conference on Quantitative Methods for Decision Making* (ORBEL 14; Mons, Jan. 20-21, 2000), pp. 32-33.

[81] D. Fiems B. Steyaert, H. Bruneel, Buffer contents in a discrete-time GI-G-1 queueing system with correlated server vacations, Book of Abstracts *Fourteenth Conference on Quantitative Methods for Decision Making* (ORBEL 14; Mons, Jan. 20-21, 2000), pp. 61-62.

[82] B. Steyaert, H. Bruneel, G.H. Petit, D. De Vleeschauwer, 'A versatile queueing model applicable in IP traffic studies', Proc. *COST 257 Management Committee Meeting* (Barcelona, Jan. 20-21, 2000).

[83] D. Fiems, B. Steyaert, H. Bruneel, 'Analysis of a discrete-time GI-G-1 queue with uncorrelated random server interruptions', Proc. *Fifth INFORMS Telecommunications Conference* (Boca Raton, March 5-8, 2000), pp. 17-18.

[84] D. De Vleeschauwer, G. Petit, S. Wittevrongel, B. Steyaert, H. Bruneel, 'An accurate closed-form formula to calculate the dejittering delay in packetised voice transport', Proc. *Networking 2000* (Paris, May 14-19, 2000), pp. 374-385.

[85] J. Walraevens, B. Steyaert, H. Bruneel, 'Analysis of packet delay in a GI-G-1 queue with non-preemptive priority scheduling', Proc. *Networking 2000* (Paris, May 14-19, 2000), pp. 433-445.

[86] V. Inghelbrecht, B. Steyaert, H. Bruneel, 'Analysis of buffers with Markov modulated server interruptions', Proc. *COST 257 Management Committee Meeting* (Kjeller, Norway, May 18-19, 2000).

[87] D. Fiems, B. Steyaert, H. Bruneel, 'Discrete-time queueing model with general service-time distribution and two-state Markovian server interruptions', Proc. *8th IFIP Workshop on Performance Modelling and Evaluation of ATM and IP Networks* (Ilkley, July 17-19 2000), pp. 25/1-25/13.

[88] J. Walraevens, B. Steyaert, H. Bruneel, 'Analysis of a GI-Geo-1 preemptive resume priority buffer', Proc. *8th IFIP Workshop on Performance Modelling and Evaluation of ATM and IP Networks* (Ilkley, July 17-19, 2000), pp. 88/1-88/11.

[89] V. Inghelbrecht, B. Steyaert, H. Bruneel, K. Laevens, 'Queueing of fixed-length messages in the presence of server interruptions', Proc. *2000 Symposium on Performance Evaluation of Computer and Telecommunication Systems* (SPECTS 2000; Vancouver, Canada, July 16-20, 2000), pp. 495-502.

[90] D. Fiems, B. Steyaert, H. Bruneel, 'Discrete-time queues with general service times and general server interruptions', Proc. *SPIE's International Symposium on Voice, Video and Data Communications* (Boston, Nov. 6-7, 2000), vol. 4211, pp. 93-104.

[91] B. Steyaert, H. Bruneel, G.H. Petit, D. De Vleeschauwer, 'Analysis of the Discrete-Time NxD-BMAP/G/1 queueing model', Proc. *International Conference on Modern Mathematical Methods of Investigating of the Information networks*, *16th Belarusian Winter Workshop on Queuing Theory* (BWWQT-01; Minsk, Jan. 23-25, 2001), pp. 36-38.

[92] D. De Vleeschauwer, A. Van Moffaert, J. Janssen, M. Büchli, G. H. Petit, B. Steyaert, H. Bruneel, 'Determining the number of packet-based phones that can be supported by one access node', Proc. *14th ITC Specialists Seminar on Access Networks and Systems* (Girona, April 25-27, 2001), pp. 197-204.

[93] J. Walraevens, B. Steyaert, H. Bruneel, 'Analysis of a preemptive resume priority buffer with general service times for the high priority class' Proc. *5th Intern. Conference on Communication Systems* (AFRICOM 2001; Cape Town, South Africa, May 28-30, 2001), pp. 1-16.

[94] D. Fiems, B. Steyaert, H. Bruneel, 'Transmission strategies for queues subjected to random interruptions', Book of Abstracts *IFIP Joint WG 6.1, 6.3 and 7.3 Workshop* (Cape Town, South Africa, May 31, 2001), paper 3.

[95] B. Steyaert, H. Bruneel, Evaluation of the packet delay in the heterogeneous N×D-BMAP/G/1 queueing system, Book of Abstracts *IFIP Joint WG 6.1, 6.3 and 7.3 Workshop* (Cape Town, South Africa, May 31, 2001), paper 13.

[96] J. Walraevens, B. Steyaert, H. Bruneel, 'Performance analysis of a GI-1-1 priority queue with a general number of priority classes', *IFIP Joint WG 6.1, 6.3 and 7.3 Workshop* (Cape Town, South Africa, May 31, 2001), paper 14.

[97] D. Fiems, B. Steyaert, H. Bruneel, 'Analysis of a discrete-time queueing model with server interruptions - modeling preemptive priority systems', Proc. *COST 279 Second Management Committee Meeting* (Lisboa, Oct. 8-9, 2001), COST279TD(01)05.

[98] J. Walraevens, B. Steyaert, H. Bruneel, 'A single-server queue with a priority scheduling discipline: performance study', Proc. *COST 279 Second Management Committee Meeting* (Lisboa, Oct. 8-9, 2001), COST279TD(01)06.

[99] D. De Vleeschauwer, A. Van Moffaert, M. Buchli, J. Janssen, G.H. Petit, B. Steyaert, H. Bruneel, 'Determining the tolerable load generated by a set of packet-based phones on a multiplexing node' Proc. *17th International Teletraffic Congress* (ITC 17; Salvador da Bahia. Brazil, Dec. 2-7, 2001), pp. 433-444.

[100] J. Walraevens, and B. Steyaert, H. Bruneel, 'Performance analysis of a GI-G-1 preemptive resume priority buffer', Proc. *Second Intern. IFIP-TC6 Networking Conference* (Networking 2002; Pisa, Italy. May 19-24. 2002), pp. 745-756.

[101] V. Inghelbrecht, B. Steyaert, S. Wittevrongel, H. Bruneel, 'Analysis of the interdeparture process in consecutive stages of a VoIP network', Proc. *COST 279 Fourth Management Committee Meeting* (Espoo, May 30-31, 2002), COST279TD(02)30.

[102] D. Fiems, B. Steyaert, H. Bruneel, 'Randomly interrupted GI-G-1 queues, service strategies and stability issues', Book of Abstracts *First Madrid Conference on Queueing Theory* (MCQT'02; Madrid, July 2-5, 2002), pp. 21-22.

[103] V. Inghelbrecht, B. Steyaert and H. Bruneel, 'Study of the burstification mechanism of an OBS edge router', Proc. *7th IFIP Working Conference on Optical Network Design & Modelling* (ONDM 2003; Budapest, Feb. 3-5, 2003), vol. II, pp. 1221-1239.

[104] J. Walraevens, and B. Steyaert, H. Bruneel, 'Analysis of a priority queue with general service times', Presentation *First Benelux Workshop on Performance Analysis of Communication Systems* (EURANDOM; Eindhoven, March 13-14, 2003).

[105] V. Inghelbrecht, B. Steyaert and H. Bruneel, 'Burstification mechanism of an OBS edge router: analytical analysis of a one- and two-threshold case', Proc. *COST 279 Management Committee Meeting* (Karlskrona, Sweden, May 22-23, 2003), COST279TD(03)37.

[106] J. Walraevens, and B. Steyaert, H. Bruneel, 'Delay analysis of a discrete-time non-preemptive priority buffer with 3 traffic classes'. Book of Abstracts *7th WSEAS International Conference on Communications* (Corfu, July 7-10, 2003), pp. 119.

[107] V. Inghelbrecht, B. Steyaert, S. Wittevrongel, H. Bruneel, 'Investigation of the interdeparture time in a multistage VoIP network', Proc. *18th International Teletraffic Congress* (ITC 18; Berlin, Aug. 31–Sep. 5, 2003), pp. 1191-1200.

[108] J. Walraevens, and B. Steyaert, H. Bruneel, 'Analysis of a preemptive repeat priority buffer with resampling', Proc. *International Network Optimization Conference*, (INOC'2003; Evry/Paris, October 27-29, 2003), pp. 581-586.

[109] J. Walraevens, and B. Steyaert, H. Bruneel, 'Priority queues with general service times', Presentation *COST 279 Mid-Term Seminar on "Analysis and Design of Advanced Multiservice Networks supporting Mobility, Multimedia, and Internetworking"* (Rome, Jan. 21-22, 2004).

[110] G. Van Hoey, D. De Vleeschauwer, B. Steyaert, V. Inghelbrecht, H. Bruneel, 'Benefit of admission control in aggregation network dimensioning for video services', Proc. *The Third IFIP-TC6 Networking Conference* (Networking 2004; Athens, May 9-14, 2004), pp. 357-368.

[111] V. Inghelbrecht, B. Steyaert, H. Bruneel, Analysis of an optical buffer with an offset-time based scheduling mechanism, Proc. *First Workshop on New Trends in Modelling, Quantitative Methods and Measurements* (Zakopane, 27-29 June 2004).

[112] J. Walraevens, B. Steyaert, M. Moeneclaey, H. Bruneel, 'A discrete-time HOL priority queue with multiple traffic classes', Proc. *of the 4th International Conference on Networking*, (ICN 2005; La Reunion, April 17-21, 2005), pp. 620-627.

[113] V. Inghelbrecht, B. Steyaert, S. Wittevrongel, H. Bruneel, 'Analysis of fiber delay line buffers with offset-time management', Presentation *Third IFIP Workshop on Next Generation Networks: Architecture, Protocols, Performance* (Placencia, Belize, March 11-12, 2005).

[114] B. Steyaert, V. Inghelbrecht, D. De Vleeschauwer, H. Bruneel, 'Some results on a versatile queueing model with heterogeneous sources', Proc. 5th *St. Petersburg Workshop on Simulation* (SIMULATION 2005; St. Petersburg, June 26 –July 2, 2005) , pp. 665-670.

[115] K. Laevens, B. Steyaert, D. De Vleeschauwer, H. Bruneel, 'Analysis and design of a play-out buffer for VBR streaming video' Book of Abstracts *Second Madrid Conference on Queueing Theory* (MCQT'06; Madrid, July 3-7, 2006), p. 38.

[116] D. Claeys, K. Laevens, B. Steyaert, H. Bruneel, 'A discrete-time queueing model with batch service', Book of Abstracts *21st National Conference of the Belgian OR Society* (ORBEL 21, Luxemburg, Jan. 18-19, 2007), pp. 96-97

[117] D. Fiems, B. Steyaert, H. Bruneel, 'Performance analysis of a partially shared buffer with correlated arrivals', Proc. *Networking 2007* (Atlanta, 14-18 May 2007).

REFERENCES

[118] J. Abate, G. L. Choudhry, W. Whitt, 'Asymptotics for steady-state tail probabilities in structured Markov queueing models', *Stochastic Models*, vol. 10(1), 1994, pp. 99–143.

[119] H. Ahmadi, W.E. Denzel, 'A survey of modern high-performance switching techniques, *IEEE Jour. on Sel. Areas in Commun.* (JSAC), vol. 7(7), 1989, pp. 1091-1103.

[120] D. Anick, D. Mitra, and M. M. Sondhi, 'Stochastic theory of a datahandling systems with multiple sources', *Bell System Tech. Jour.*, vol. 61(8), Oct. 1982, pp. 1871-1894.

[121] N. Ansari, H. Liu, Q. Shi, 'On modeling MPEG video traffics', *IEEE Trans. on Broadcasting*, vol.48(4), Dec. 2002, pp. 337-347.

[122] R.Y. Awdeh, H.T. Mouftah, 'Survey of ATM switch architectures', *Computer Networks and ISDN systems*, vol. 27, 1995, pp. 1567-1613.

[123] S. Baey, 'Modeling MPEG-4 video traffic based on a customization of the DBMAP', Proc. *2004 International Symposium on Performance Evaluation of Computer and Telecommunication Systems* (SPECTS'04; San Jose, CA, July 25-29, 2004), pp. 705-714.

[124] A. Baiocchi, 'Asymptotic behaviour of the loss probability of the $M/G/1/K$ and $G/M/1/K$ queues', *Queueing Systems*, vol. 10(3), Jan. 1992, 235-248.

[125] A. Baiocchi, 'Analysis of the loss probability of the $MAP/G/1/K$ queue. Part 1 : asymptotic theory', *Commun. Statist. – Stochastic Models*, vol. 10(4), 1994, 867-893.

[126] G. P. Basharin, A.N. Langville, V.A. Naumov, 'The life and work of A.A. Markov', *Jour. Linear Algebra and its Applications*, vol. 386, July 2004, pp. 3-26.

[127] B. Bensaou, J. Guibert, J.W. Roberts, 'Fluid queueing models for a superposition of ON/OFF sources', Proc. *7th ITC Specialist Seminar* (Morristown, NJ, Oct. 1990), pp. 13-19.

[128] C. Bisdikian, J.S. Lew and A.N. Tantawi, 'On the tail approximation of the blocking probability of single server queues with finite buffer capacity', Proc. *Second International Conference on Queueing Networks with Finite Capacity* (Research Triangle Park, NC, May 28-29, 1992), pp. 267-280.

[129] C. Blondia, O. Casals, 'Statistical multiplexing of VBR sources: a matrix analytic approach', *Performance Evaluation*, vol. 16, 1992, pp. 5-20.

[130] C. Blondia, 'A discrete time batch Markovian arrival process as B-ISDN traffic model', *Belgian Journal of Operations Research Statistics and Computer Science*, vol. 32, 1993, pp. 3–23.

[131] E. Brandt, 'De wiskunde van de ergernis', *De Groene Amsterdammer*, Jan. 1996.

[132] U. Briem, T.H. Theimer, H. Kröner, 'A general discrete-time queueing model: analysis and applications', Proc. *13th International Teletraffic Congress* (ITC-13; Copenhagen, June 1991), pp. 13-19.

[133] P. Brown, S. Simonian, 'Perturbation of a periodic flow in a synchronous server', Proc. *Performance `87* (Brussels, 1987), pp. 89-112.

[134] A. Bruin, A. Rossum, M. Visser, G. Koole, 'Modeling the emergency cardiac in-patient flow: an application of queuing theory' *Health Care Management Science*, vol. 10 (2), 2007, pp. 125-137.

[135] H. Bruneel, 'Message delay in TDMA channels with contiguous output', *IEEE Trans. on Commun.*, vol. 34, 1986, pp. 681-684.

[136] H. Bruneel, B.G. Kim, 'Discrete-time models for communication systems, including ATM', *Kluwer Academic publishers* (ISBN: 0-7923-9292-2), 1992.

[137] H. Bruneel, S. Wittevrongel, 'An approximate analytical technique for the performance evaluation of ATM switching elements with burst routing', *Computer Networks and ISDN Systems*, vol. 28, 1996, pp. 325-343.

[138] J. A. Buzacott, D. D. Yao, 'On queueing network models of flexible manufacturing systems', *Queueing Systems: Theory and Applications*, vol. 1(1), 1986, pp. 5-27.

[139] C.G. Chang, H.H. Tan, 'Queueing analysis of explicit policy assignment push-out buffer sharing schemes for ATM networks', Proc. *IEEE Conference on Computer Communications* (INFOCOM'94, Toronto, June 12-16 1994), pp. 500-509.

[140] G.L. Chaudry, D.M. Lucantoni, W. Whitt, 'Squeezing the most out of ATM' *IEEE Trans. on Commun.*, vol. 44(2), Feb. 1996, pp. 203-217.

[141] J.S-C. Chen, R. Guérin, T.E. Stern, 'Markov-modulated flow model for the output queue of a packet switch', *IEEE Trans. on Commun*., vol. 40(6), June 1992, pp. 1098-1110.

[142] I. Cidon, R. Guérin, A. Khamisy, 'On protective buffer policies' *IEEE Trans. on Networking*, vol. 2(3), June 1994, pp. 240-246.

[143] I. Cidon, L. Georgiadis, R. Guérin, A. Khamisy, 'Optimal buffer sharing', *IEEE Jour. on Sel. Areas in Commun.*, vol. 13(7), Sep. 1995, pp. 1229-1240.

[144] J.W. Cohen, 'On the asymmetric clocked buffered switch', *Queueing Systems: Theory and Applications*, vol. 30(3-4), 1998, 385-404.

[145] J.H. Cozzens, L.A. Finkelstein, 'Range and error analysis for a fast Fourier transform computed over $Z[\omega]$', *IEEE Trans. on Information Theory*, vol. 33(4), July 1987, pp. 582-590.

[146] C.D. Crommelin, 'Delay probability formulae when the holding times are constant', *Post 0ffice Electrical Engineers Journal*, vol. 25, 1932, pp. 41-50.

[147] E. Del Re and R. Fantacci, 'Performance evaluation of input- and output queueing techniques in ATM switching systems', *IEEE Trans. on Commun.*, vol. 41, 1993, pp. 1565-1575.

[148] D. Denteneer, J.S.H. van Leeuwarden, J.A.C. Resing, 'Bounds for a discrete-time multiserver queuing model with an application to cable networks', Proc. *18th International Teletraffic Congress*, (ITC18; Berlin, Aug. 31- Sep. 1, 2003).

[149] M. De Prycker, 'Asynchronous Transfer Mode – solutions for broadband ISDN' *Prentice-Hall*, 3rd edition, 1995.

[150] D. De Vleeschauwer, J. Janssen, G.H. Petit, 'Voice over IP in access networks', Proc. *7th IFIP Workshop on Performance Modelling and Evaluation of ATM/IP Networks* (IFIP ATM '99; Antwerp, Be., June 28-30, 1999).

[151] D. De Vleeschauwer, J. Janssen, G. H. Petit, F. Poppe, 'Quality bounds for packetized voice transport', *Alcatel Telecom Review*, First quarter, Jan. 2000, pp. 19-23.

[152] A. Doig, 'A bibliography on the theory of queues', *Biometrika*, vol. 44 (34), 1957, pp. 490-514.

[153] P.L. Douillet, A.-L. Beylot, M. Becker, 'Computing stochastical bounds and tail distribution of an *M*/*GI*/1 Queue', *Lecture Notes In Computer Science*; vol. 1815, 2000, pp. 423-432.

[154] J. Dshalalow, 'Frontiers in queueing: models and applications in science and engineering', *Probability and Stochastic Series*, *CRC Press Inc*., (Boca Raton, ISBN:0-8493-8076-6), 1997.

[155] K.Y. Eng, M.J. Karol, Y.-S. Yeh, 'A growable packet (ATM) switch architecture: design principles and application', *IEEE Trans. on Commun.*, vol. 40(2), Feb. 1992, pp. 423-430.

[156] R. Epsilon, J. Ke, and C. Williamson, 'Analysis of ISP IP/ATM Network Traffic Measurements', *ACM Sigmetrics Performance Evaluation Review*, vol. 27(2), Sept. 1999, pp. 15-24.

[157] A.K. Erlang, 'Calcul des probabilités et conversations téléphoniques', *Revue Générale de L'électricité*, vol. XVIII(8), Aug. 1925, pp. 305-309.

[158] A.E. Ferdinand, 'An analysis of the machine interference model', *IBM Syst. Jour.*, no. 2, 1971, pp. 129-142.

[159] D. Ferrari, D. Verma, 'A scheme for real-time channel establishment in wide-area networks', *IEEE Jour. on Sel. Areas in Commun.*, vol. 8 (3), Apr. 1990, pp. 368-379.

[160] D. Fiems, H. Bruneel, 'A note on the discretization of Little's result', *Operations Research Letters*, vol. 30, 2002, pp. 17-18.

[161] W. Fischer, K. Meier-Hellstern, 'The Markov-modulated Poisson process (MMPP) cookbook', *Performance Evaluation*, vol. 18(2), Sep. 1992, pp. 149-171.

[162] S. Floyd, V. Jacobson, 'Random early detection gateways for congestion avoidance', *IEEE/ACM Trans. on Networking*, vol. 1(4), Aug. 1993, 197-413.

[163] H.R. Gail, S.L. Hantler, A.G. Konheim, B.A. Taylor, 'An analysis of a class of tele-communication models', *Performance Evaluation*, vol. 21, 1994, pp. 151-161.

[164] R.G. Gallager, 'Discrete stochastic processes', *Kluwer Academic publishers* (ISBN: 0-7923-9292-2), 1996.

[165] P. Gao, S. Wittevrongel, H. Bruneel, 'Discrete-time multiserver queues with geometric service times', *Computers and Operations Research*, vol. 31(1), Jan. 1994, pp. 81-99.

[166] J. García, O. Casals, 'Stochastic models of space priority mechanisms with Markovian arrival processes', *Annals of Operations Research*, vol. 35, Jan. 1992, pp. 271-296.

[167] D.C. Gazis, 'The origins of traffic theory', *Operations Research*, vol. 50(1), Feb. 2002, pp. 69-77.

[168] F. Geerts and C. Blondia, 'Superposition of Markovian sources and long range dependence', Proc. *IFIP TC6/WG6.2 Fourth International Conference on Broadband Communications* (BC '98; Stuttgart, Apr. 1-2, 1998), pp. 550-561.

[169] P.W. Glynn, W. Whitt, 'Logarithmic asymptotics for steady-state tail probabilities in a single-server queue'*, Journal of Applied Probability*, vol. 31, 1994, pp. 131-156.

[170] F.N. Gouweleeuw, H.C. Tijms, 'Computing loss probabilities in discrete-time queues', *Operations Research*, vol. 46(1), Jan. 1998, pp. 149-154.

[171] A. Gravey, G. Hébuterne, 'Simultaneity in discrete-time single server queues with Bernoulli inputs', *Performance Evaluation*, vol. 14(2), 1992, pp. 123-131.

[172] A. Guerrero and F. Lozano, 'Queuing models applying to a class of ATM multiplexers', Proc. *13th International Teletraffic Congress* (ITC-13; Copenhagen, June 1991), pp. 999-1004.

[173] F. Guillemin, J. Boyer, A. Dupuis, 'Burstiness in broadband integrated networks, *Performance Evaluation*, vol. 15, 1992, 163-176.

[174] M. Hamdi, J. Muppala, 'Performance evaluation of ATM switches under various traffic and buffering schemes,' Proc. *IEEE International Global Telecommunications Conference* (GLOBECOM'95, Singapore, Nov. 13-17, 1995), pp. 828-832.

[175] P. Harrison, 'Performance engineering and stochastic modelling'*, Lecture Notes in Computer Science*, vol. 3670, Sept. 2005, pp.1–14.

[176] O. Hashida, Y. Takahashi and S. Shimogara, Switched batch Bernoulli process (SBBP) queue with application to statistical multiplexer performance, *IEEE Jour. on Sel. Areas in Commun.*, vol. 9(3), 1991, pp. 394–401.

[177] G. Hébuterne, A. Gravey, 'A space priority queuing mechanism for multiplexing ATM channels', *Computer Networks and ISDN Systems*, vol. 20, 1993, pp. 37-43.

[178] G. Hébuterne, A. Dupuis, 'Microscopic models of ATM multiplexing', Proc. *6th FIP Workshop on Performance Modelling and Evaluation of ATM/IP Networks* (Ilkley, UK, July 8-10, 1996), pp. 55/1-9.

[179] H. Heffes, D.M. Lucantoni, 'A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE Jour. on Sel. Areas in Commun.*, vol. 4(6), 1986, pp. 856-867.

[180] M.A. Henrion, G.J. Eilenberger, G.H. Petit, P.H. Parmentier, 'A multipath self-routing switch', *IEEE Communications Magazine*, vol. 31(4), Apr. 1993, pp. 46-52.

[181] C. Hermann, 'The complete analysis of the DBMAP/G/1/*N* queue', *Performance Evaluation*, vol. 43, 2001, pp. 95-121.

[182] M.G. Hluchyj, M.J. Karol, 'Queueing in high-performance packet switching', *IEEE Jour. on Sel. Areas in Commun.* (JSAC), vol. 6, no. 9, Dec. 1988, pp. 1587-1597.

[183] J.Y. Hui, E. Arthurs, 'A broadband packet switch for integrated transport', *IEEE Jour. on Sel. Areas in Commun.* (JSAC), vol. 5(8), 1987, pp. 1264-1273.

[184] G.U. Hwang, B.D. Choi, 'Closed-form expressions on the geometric tail behavior of statistical multiplexers with heterogeneous traffic', *IEEE Trans. on Commun.*, vol. 46(12), Dec. 1998, pp. 1575-1579.

[185] F. Ishizaki, T. Takine, 'Cell loss probability approximation and their application to call admission control' *Adv. Performance Anal.*, vol. 2**,** 1999, pp. 225-258.

[186] ITU-T, 'Objective measurements of active speech level', *ITU-T Recommendation P.56*, March 1993.

[187] J. Janssen, D. De Vleeschauwer, G.H. Petit, 'Delay and distortion bounds for packetized voice calls of traditional PSTN quality', Proc. *First IP Telephony Workshop* (IPTel 2000; Berlin, April 12-13, 2000), GMD Report 95, pp. 105-110.

[188] A. Jeffrey, 'Complex analysis and its applications', *CRC Press* (ISBN: 0-8493-8623-3), 1992.

[189] M.G. Kienzle, K.C. Sevcik, 'Survey of analytic queueing network models of computer systems', *ACM Sigmetrics Performance Evaluation Review*, vol. 8(3), 1979, pp. 113-129.

[190] M. Karol, M. Hluchyj, S. Morgan, 'Input versus output queueing in a space-division packet switch', *IEEE Trans. on Commun.*, vol. 35(12), 1987, pp.1347-1355.

[191] M. Katevenis, S. Sidiropoulos, C. Courcoubetis, 'Weighted round-robin cell multiplexing in a general-purpose ATM switch chip', *IEEE Jour. on Sel. Areas in Commun.*, vol. 9(8), Oct. 1991, pp. 1265-1279.

[192] H.S. Kim, N.B. Schroff, 'Loss probability calculations and asymptotic analysis for finite buffer multiplexers', *IEEE/ACM Trans. on Networking*, vol. 9(6), Dec. 2001, pp. 755-768.

[193] H.S. Kim, N.B. Schroff, 'On the asymptotic relationship between the overflow probability and the loss ratio', *Advances in Applied Probability*, vol. 33(4), 1999, pp. 836-863.

[194] B. Kim, J. Kim, I-S. Wee, B.D. Choi, 'Asymptotic analysis of loss probability in a finite queue where one packet occupies as many places as its length', *Performance Evaluation*, vol. 54, 2003, pp. 209-223.

[195] L. Kleinrock, 'Queueing Systems, Volume I: Theory', *John Wiley & Sons* (ISBN: 978-0-471-49110-1), 1975.

[196] H. Kröner, G. Hébuterne, P. Boyer, A. Gravey, 'Priority management in ATM switching nodes', *IEEE Jour. on Sel. Areas in Commun.*, vol. 9(3), Apr. 1991, pp. 418-427.

[197] C.P. Kruskal, M. Snir, A. Weiss, 'The distribution of waiting times in clocked multistage interconnection networks', *IEEE Trans. on Comp.*, vol. 37(11), Nov. 1988, pp. 1337-1353.

[198] R.H.Kwong, A.L. Garcia, 'Performance analysis of an integrated hybrid-switched multiplex structure', *Performance Evaluation*, vol. 4, 1984, pp. 81-91.

[199] M.A. Labrador, S. Banerjee, 'Packet dropping policies for ATM and IP Networks', *IEEE Communications Surveys*, vol. 2(3), 1999, pp. 2-14.

[200] K. Laevens, H. Bruneel, 'Delay analysis for ATM queues with random order of service', *Electronics Letters*, vol. 31(5), 1995, pp. 346-347.

[201] K. Laevens, 'Stochastische modellering van ATM-schakelelementen met buffers aan de ingangszijde', *Proefschrift ingediend tot het behalen van de graad van doctor in de toegepaste wetenschappen*, 1999.

[202] R,C. Larson, 'Perspective on queues : social justice and the psychology of queueing', *Operations Research*, vol. 35 (6), Dec. 1987, pp. 895-905.

[203] G. Latouche, V. Ramaswami, 'Introduction to matrix analytic methods in stochastic modeling', *Cambridge University Press* (ISBN: 0-8987-1425-7), 2002.

[204] J.-Y. Le Boudec, 'An efficient solution method for Markov Models of ATM links with loss priorities', *IEEE Jour. on Sel. Areas in Commun.*, vol. 9(3), Apr. 1991, pp. 408-416.

[205] T.T. Lee, 'A modular architecture for very large packet switches', *IEEE Trans. on Commun.*, vol. 38(7), 1990, pp. 1097-1106.

[206] H.W. Lee, J.M. Moon, B.K. Kim, J.G. Park, S.W. Lee, 'A simple eigenvalue method for low-order D-BMAP/G/1 queues', *Applied Mathematical Modelling*, Volume 29(3), Mar. 2005, Pages 277-288.

[207] S. Li, H. Sheng, 'Discrete queueing analysis of multimedia traffic with diversity of correlation and burstiness properties', Proc. *IEEE Conference on Computer Communications* (INFOCOM `91; Miami, April 7-11, 1991), pp. 368-381.

[208] S.-Q. Li, J.W. Mark, 'Performance of voice/data integration on a TDM system', *IEEE Trans. on Commun*., vol. 33(12), Dec. 1985, pp. 1265-1273.

[209] S.-Q. Li, 'Study of information loss in packet voice systems', *IEEE Trans. on Commun*., vol. 37, Nov. 1989, pp. 1192-1202.

[210] S.-Q. Li, 'A general solution technique for discrete queueing analysis of multimedia traffic on ATM', *IEEE Trans. on Commun.*, vol. 39(7), 1991, pp. 1115-1132.

[211] S.Q. Li, 'Discrete queueing analysis of multimedia traffic with diversity of correlation and burstiness properties', *IEEE Trans. on Commun.*, vol. 42(2-4), Feb. 1994, pp. 1339-1351.

[212] A.Y.-M. Lin, J.A. Silvester, 'Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system', *IEEE Jour. on Sel Areas in Commun.*, vol. 9(9), Dec. 1991, pp. 1524-1536.

[213] D.V. Lindley, 'The theory of queues with a single server', *Proceedings of the Cambridge Philosophical Society*, vol. 48, 1952, pp. 277-289.

[214] M. Listani, A. Roveri, 'Switching structures for ATM', *Computer Communications*, vol. 12(6), 1989, pp. 349-358.

[215] J. Little, 'A proof for the queueing formula: $L=\lambda W$', *Operations Research*, vol. 9, 1961, pp. 383-397.

[216] K. Liu, D.W. Petr, V.S. Frost, H. Zhu, 'Design and analysis of a bandwidth management framework for ATM-based broadband ISDN', *IEEE Communications Magazine*, May 1997, pp. 138-145.

[217] J.R. Louvion, P. Boyer, A. Gravey, 'A discrete-time single server queue with Bernoulli arrivals and constant service time', Proc. *12th International Teletraffic Congress* (ITC13; Torino, June 1-8, 1988), pp. 1304-1311.

[218] D.M. Lucantoni, 'New results on the single server queue with a batch Markovian arrival process', *Stochastic Models*, vol. 7(1), 1991, pp. 1–46.

[219] R.J. Mc Millen, 'A survey of interconnection networks', Proc. *IEEE Global Telecommunications Conference*, (GLOBECOM `84; Atlanta, GA, Nov. 26-29, 1984), pp. 105-113.

[220] R. Mahnke, R. Kaupuz, 'Probabilistic description of traffic flow', *Networks and Spatial Economics*, vol. 1(1-2), 2001, pp. 103-136.

[221] T. Meisling, 'Discrete-time queuing theory', *Operations Research*, vol. 6(1), Jan. 1958, pp. 96-105.

[222] H. Michiel, K. Laevens, 'Teletraffic engineering in a broad-band era', *Proc. of the IEEE*, vol. 85(12), Dec. 1997, pp. 2007-2033.

[223] S. Minzer, 'Broadband ISDN and Asynchronous Transfer Mode (ATM)', *IEEE Communications Magazine*, vol. 20(9), 1989, pp. 17-24.

[224] D. Moltchanov, Y. Koucheryavy, J. Harju, 'The model of single smoothed MPEG traffic source based on the D-MAP arrival process with limited state space', Proc. *5th International Conference on Advanced Communication Technology* (ICACT'2003; Gangwon-Do, South Korea, Jan. 20-22, 2003), pp. 57-63.

[225] E. Morozov (1999). 'Self-similarity and long-range dependence in network traffic modeling', Proc. *Developments in Distributed Systems and Data Communications* (FDPW'99; Petrozavodsk, 1999), pp. 32-45.

[226] A. Myskja, 'An Introduction to Teletraffic', *Telektronikk*, vol. 91, No. 2/3, pp. 3-41, 1995.

[227] M.F. Neuts, 'Matrix-geometric solutions in stochastic models: an algorithmic approach', *Johns Hopkins Univ. Press* (ISBN: 0-4866-8342-7), Baltimore, 1981.

[228] M.F. Neuts, 'Structured stochastic matrices of *M*/*G*/1 type and their applications', *Marcel Dekker Inc* (ISBN: 0-8247-8283-6), New York, 1989.

[229] C.H. Ng, L. Bai, B.H. Soong, 'Modelling multimedia traffic over ATM using MMBP' *IEE Proceedings Communications*, vol. 144(5), Oct. 1997, pp. 307-310.

[230] I. Norros, J, W. Roberts, A. Simonian, and I. T, Virtamo, 'The superposition of variable bit rate sources in an ATM multiplexer', *IEEE Jour. on Sel. Areas in Commun.* (JSAC), vol. 9, Apr. 1991, pp. 378-387.

[231] Y. Oie, M. Murata, K. Kubota, H. Miyahara, 'Performance analysis of nonblocking packet switch with input and output buffers', *IEEE Trans. on Commun*., vol. 40(8), Aug. 1992, pp. 1294-1297.

[232] A. Pattavina, 'An ATM switch architecture for provision of integrated broadband services, *IEEE Jour. on Sel. Areas in Commun.* (JSAC), vol. 9(9), 1991, pp. 1537-1548.

[233] C. Palm, 'Methods of judging the annoyance caused by congestion', *TELE*, vol. 2, 1953, pp. 1-20.

[234] A.K. Parekh, R.G. Gallager, 'A generalized processor sharing approach to flow control in integrated services networks : the single-node case', *IEEE/ACM Trans. on Networking, vol. 1(3), June 193, pp. 344-357.*

[235] A. Pattavina, 'Nonblocking architectures for ATM switching', *IEEE Communications Magazine*, Feb. 1993, pp. 38-48.

[236] L.K. Platzmann, J.C. Ammons, J.J. Bartholdi, 'A simple and efficient algorithm to compute tail probabilities from transforms', *Oprations Research*, vol. 36(1), Jan. 1988, pp. 137-144.

[237] N. U. Prabhu, 'A bibliography of books and survey papers on queueing systems: theory and applications', *Queueing Systems : Theory and Applications*, vol. 2(4), Feb. 1988, pp. 393-398.

[238] J. Preater, 'A bibliography of queues in health and medicine', *Keele Mathematics Research Report*, Issue 01-1, 2001, pp. 1-11.

[239] M. Rahnema, 'Frame relay and the fast packet switching concepts and issues', *IEEE Network Magazine*, vol. 5(4), July 1991, pp. 18-23.

[240] R. Ramdjee, J. Kurose, D. Towsley, H. Schulzrinne, 'Adaptive playout mechanisms for packetized audio applications in wide-area networks', Proc. *IEEE Conference on Computer Communications* (INFOCOM 94; Toronto, June 12-16, 1994), pp. 680-688.

[241] T.L. Saaty, 'Résumé of useful formulas in queueing theory', *The Journal of the Operations Research Society of America*, vol. 5(2), Apr. 1957, pp. 161-200.

[242] T. Shioyama, H. Kise, 'Optimization in production systems – a survey on queuing approaches', *Journal of the Operations Research Society of Japan*, vol. 32(1), Mar. 1989, pp. 34-55.

[243] K. Sohraby, 'Heavy traffic multiplexing behavior of heterogeneous bursty sources and their admission control in ATM networks', Proc. *IEEE Global Telecommunications Conference*, (GLOBECOM '92; Orlando, FL, Dec. 6-9, 1992), pp. 1518-1523.

[244] K. Sohraby. 'On the theory of general ON/OFF sources with applications in high speed networks', Proc. *IEEE Conference on Computer Communications*, (INFOCOM 93; San Francisco, CA, Mar. 30-Apr. 1, 1993), pp. 401-410.

[245] T. E. Stem and A. I. Elwalid, 'Analysis of separable Markov-modulated rate models for information-handling systems,' *Advances in Applied Probability*, vol. 23, 1991, pp. 105-139.

[246] S. Stidham, 'Analysis, design and control of queueing systems', *Operations Research*, vol. 50(1), Jan. 2002, pp. 197-216.

[247] K. Stordahl, 'The history behind the probability theory and the queueing theory', *Telektronikk* 2, 2007, pp. 123-140.

[248] H. Takagi, 'Queueing analysis, volume 3 : discrete-time systems' *North Holland*, Amsterdam (ISBN: 0-4448-1611-9) 1993.

[249] H. Takagi, L.B. Boguslavsky, 'A supplementary bibliography of books on queueing analysis and performance evaluation', *Queueing Systems : Theory and Applications*, vol. 8(3), Apr. 1991, pp. 313-322.

[250] F. Tobagi, 'Fast packet switch architectures for broadband integrated services digital networks', *Proceedings of the. IEEE*, vol. 78(1), 1990, pp. 133-165.

[251] T. Tsuchiya, Y. Takahashi, 'On discrete-time single-server queues with Markov modulated batch Bernoulli input and finite capacity', *Jour. of the Oper. Soc*., vol. 36(1), March 1993, pp. 29-45.

[252] J. Turner, 'Design of an integrated services packet network' *IEEE Jour. on Sel. Areas in Commun.* (JSAC), vol. 4(8), 1986, pp.1373-1380.

[253] R. Ulrich, U. Herzog, P. Kritzinger, 'Modeling buffer utilization in cell-based networks', *Performance Evaluation*, vol. 31, 1998, pp. 183-199.

[254] F.A. Van der Duyn Schouten, S.G. Vanneste, 'Maintenance optimization of a production system with buffer capacity', *Eur. Jour. of Oper. Res.*, vol. 82, 1995, pp. 323-338.

[255] K. Van Der Wal, M. Mandjes, H. Bastiaansen, 'Delay Performance Analysis of the New Internet Services with Guaranteed QoS', *Proceedings of the IEEE*, vol. 85(12), Dec. 1997, pp. 1947-1957.

[256] N.M. Van Dijk, **'**Why queuing never vanishes', *European Journal of Operational Research*, vol. 99 (2), June 1997, pp. 463-476.

[257] P. Van Mieghem, 'The asymptotic behavior of queueing systems: Large deviations theory and dominant pole approximation', *Queueing Systems*, vol. 23(1-4), March 1996, pp. 27-55.

[258] J.C.W. Van Ommeren, 'The discrete-time single-server queue', *Queueing Systems*, vol. (8), 1991, pp. 279-294.

[259] B. Vinck, H. Bruneel, 'General relationships between queue length and delay in discrete-time single-server queues', Proc*. of the COST 257 Management Committee Meeting* (Espoo, Finland, 27-28 May 1997).

[260] A.D. Wall, D.J. Worthington, 'Using discrete distributions to approximate general service time distributions in queueing models' *The Journal of the Operational Research Society*, vol. 45(12), Dec. 1994, pp. 1398-1404.

[261] S. Wittevrongel, H. Bruneel, 'Exact calculation of buffer-contents variance and delay jitter in a discrete-time queue with correlated input traffic', *Electronics Letters*, vol. 32(14), 1996, pp. 1258-1259.

[262] C.M. Woodside, E.D.S. Ho, 'Engineering calculation of overflow probabilities in buffers with Markov-interrupted service', *IEEE Trans. on Commun.*, vol. 35(12), Dec. 1987, pp. 1272-1277.

[263] K. Wuyts, R. Boel, 'A matrix geometric algorithm for finite buffer systems, with B-ISDN applications', *10th ITC Specialist Seminar on Control in Communications* (Lund, Sept 17-19, 1996).

[264] Y. Xiong, H. Bruneel, **'**A tight upper bound for the tail distribution of the buffer contents in statistical multiplexers with heterogeneous MMBP traffic sources', Proc. *IEEE Global Telecommunications Conference* (GLOBECOM '93; Houston, Nov. 29-Dec. 2, 1993), vol. 2, pp. 767-771.

[265] Y. Xiong, H. Bruneel, 'An analytic approach to obtain tail distributions of buffer contents and delay in a discrete-time single-server queue with bursty arrivals', *Journal of the Belgian Operations Research Society* (JORBEL), 1994, vol. 34(4), pp. 3-13.

[266] Y. Xiong, H. Bruneel, 'A simple approach to obtain tight upper bounds for the asymptotic queueing behavior of statistical multiplexers with heterogeneous traffic', *Performance Evaluation*, vol. 2, 1995, pp. 159-173.

[267] N. Yin, M.G. Hluchyj, 'Implications of dropping packets from the front of a queue', *IEEE Trans. on Commun.*, vol. 41(6), June 1993, pp. 846-851.

[268] E.W. Zegura. 'Architectures for ATM switching systems', *IEEE Communications Magazine*, Feb. 1993, pp. 28-37.

[269] L. Zhang, 'VirtualClock : a new traffic control algorithm for packet-switched networks', *ACM Trans. on Comp. Sys.*, vol. 9(2), May 1991, pp. 101-124.

[270] Z. Zhang, 'Analysis of discrete-time queue with integrated bursty inputs in ATM networks', *Intern. J. on Digital and Analog Commun. Syst.*, vol. 4, 1991, pp. 191-203.

[271] Z. Zhang, A. Acampora, 'Effect of on/off distributions on the cell loss probability in ATM networks', Proc. *IEEE Global Telecommunications Conference*, (GLOBECOM '92; Orlando, FL, Dec. 6-9, 1992), pp. 1533-1540.

[272] Z. Zhang, A.S. Acampora, 'Equivalent bandwidth for heterogeneous sources in ATM networks', Proc. *International Conference on Communications* (ICC '94; Orlando, Fl., May 1-5, 1994), pp. 1025-1031.

[273] J-A. Zhao, B. Li, I. Ahmad, 'Traffic modeling for layered video', *IEEE International Conference on Multimedia and Expo* (ICME '03; Baltimore, July 2003), pp. 497-500.

[274] J-A. Zhao, B. Li, X.-R. Cao, I. Ahmad, 'A matrix-analytic solution for the *D-BMAP/PH/*1 priority queue', *Queueing Systems*, vol. 53(3), July 2006, pp. 127-145