Discrete-tijd-wachtlijnmodellen met prioriteiten

Discrete-time queueing models with priorities

Joris Walraevens

Promotor: Prof. dr. ir. H. Bruneel

Proefschrift ingediend tot het behalen van de graad van
Doctor in de Toegepaste Wetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: Prof. dr. ir. H. Bruneel
Faculteit Toegepaste Wetenschappen
Academiejaar 2004–2005

UNIVERSITEIT GENT

Dankwoord

Hierbij wens ik iedereen te bedanken die bij mijn doctoraatsonderzoek en/of het schrijven van deze thesis (van dichtbij of van iets verderaf) betrokken was.

Eerst en vooral wens ik mijn promotor Herwig Bruneel te bedanken voor de begeesterende manier waarop hij mij aangezet heeft om zelfstandig onderzoek te voeren en specifiek om mij te verdiepen in het mooie domein van de wachtlijntheorie. Ook bedankt om naast de heel belangrijke morele steun ook steeds te zorgen voor de even belangrijke financiële ondersteuning.

Verder wil ik alle collega's en ex-collega's van vakgroep TELIN (en in het bijzonder diegenen van de SMACS onderzoeksgroep) bedanken voor de aangename werksfeer. Alhoewel ik iedereen wel met naam zou willen noemen, moet ik mij beperken tot een aantal. Annette Nevejans bedank ik voor de warme manier waarop ze steeds klaarstaat om de praktische beslommeringen in de vakgroep te regelen. Bart Steyaert en Sabine Wittevrongel bedank ik om mij bij te schaven in het schrijven van wetenschappelijke artikels en om steeds te antwoorden op allerlei vragen (van wetenschappelijke of iets minder wetenschappelijke aard). Tenslotte wil ik mijn bureaugenoten door de jaren heen (en met name de 'constanten' Dieter Fiems en Stijn De Vuyst) bedanken voor de levendige en produktieve discussies.

Tenslotte bedank ik mijn ouders, familie en vrienden voor de blijvende aanmoedigingen (en dit vooral gedurende de laatste maanden van het schrijven van dit boekje).

Bedankt, Joris.

Contents

List of Figures

Notations, acronyms, ...

ATM	Asynchronous Transfer Mode
a_j	number of class-j arrivals during a random slot
$a_{j,k}$	number of class-j arrivals during slot k
a_T	total number of arrivals during a random slot
$a_{T,k}$	total number of arrivals during slot k
$A(z_1, z_2)$	joint pgf of the number of per-slot class-j ($j = 1, 2$) arrivals
$A_j(z)$	pgf of the number of per-slot class-j arrivals
$A_T(z)$	pgf of the total number of per-slot arrivals
α	the fraction of class-1 load in the overall traffic load
$B(z)$	the pgf of a busy period
Corr[...,...]	the correlation coefficient between two stochastic variables
Cov[...,...]	the covariance of two stochastic variables
d	delay of a random unit in steady-state
d_j	delay of a random class-j unit in steady-state
$D(z)$	pgf of the delay of a random unit in steady-state
$D_j(z)$	pgf of the delay of a random class-j unit in steady-state
E[...]	the expected value operator
$E_j(z_1, z_2)$	joint pgf of the number of class-1 and class-2 arrivals during the service time of a class-j unit
$f_{1,k}^{(1)}$	the number of class-1 arrivals during the arrival-slot of a tagged class-1 unit (slot k by definition), served before this tagged unit
$f_{j,k}^{(2)}$	the number of class-j arrivals during the arrival-slot of a tagged class-2 unit (slot k by definition), served before this tagged unit
$f_{T,k}^{(2)}$	the total number of arrivals during the arrival-slot of a tagged class-2 unit (slot k), served before this tagged unit
$F_1^{(1)}(z)$	pgf of the number of class-1 arrivals during the arrival-slot of a tagged class-1 unit, served before this tagged unit
$F^{(2)}(z_1, z_2)$	joint pgf of the number of class-1 and class-2 arrivals during the arrival-slot of a tagged class-2 unit, served before this tagged unit

$F_T^{(2)}(z)$	pgf of the total number of arrivals during the arrival-slot of a tagged class-2 unit, served before this tagged unit
iff	if and only if
IP	Internet Protocol
$I(z)$	pgf of an idle period
λ_j	arrival rate of class-j units
λ_T	total arrival rate
μ_j	mean service time of a class-j unit
NP	Non-Preemptive
pgf	probability generating function
pmf	probability mass function
PR	Preemptive Resume
PRD	Preemptive Repeat Different
PRI	Preemptive Repeat Identical
Prob[...]	the probability operator
QoS	Quality of Service
$q_{j,k}$	queue contents of class-j at the beginning of slot k
$Q(z_1, z_2)$	joint pgf of the steady-state queue contents of class-1 and class-2 at the beginning of a random slot
ρ_j	arrival load of class-j
ρ_T	total arrival load
$\rho_{u_1 u_2}$	the correlation coefficient of the steady-state class-1 and class-2 system contents at the beginning of a random slot
s_j	service time of a class-j unit
$S_j(z)$	pgf of the service time of a class-j unit
t	waiting time of a random unit in steady-state
t_j	waiting time of a random class-j unit in steady-state
$T(z)$	pgf of the waiting time of a random unit in steady-state
$T_j(z)$	pgf of the waiting time of a random class-j unit in steady-state
u_j	steady-state system contents of class-j at the beginning of a random slot
$u_{j,k}$	system contents of class-j at the beginning of slot k
u_T	total steady-state system contents at the beginning of a random slot
$u_{T,k}$	total system contents at the beginning of slot k
$U_k(z_1, z_2)$	joint pgf of the system contents of class-1 and class-2 at the beginning of slot k
$U(z_1, z_2)$	joint pgf of the steady-state system contents of class-1 and class-2 at the beginning of a random slot
$U_j(z)$	pgf of the steady-state system contents of class-j at the beginning of a random slot
$U_T(z)$	pgf of the total steady-state system contents at the beginning of a random slot
v	sub-busy period initiated by a unit

v_j	sub-busy initiated by a class-j unit
$V(z)$	pgf of a sub-busy period initiated by a unit
$V_j(z)$	pgf of a sub-busy period initiated by a class-j unit
w_j	steady-state unfinished work of class-j at the beginning of a random slot
$W(z_1, z_2)$	joint pgf of the steady-state unfinished work of class-1 and class-2 at the beginning of a random slot
y	number of class-2 arrivals during a sub-busy period initiated by a random unit
y_j	number of class-2 arrivals during a sub-busy period initiated by a class-j unit
$Y(z)$	pgf of the number of class-2 arrivals during a sub-busy period initiated by a unit
$Y_j(z)$	pgf of the number of class-2 arrivals during a sub-busy period initiated by a class-j unit

Samenvatting

S.1 Inleiding

S.1.1 Wachtlijnen en buffers

Wachtlijnen maken deel uit van het dagelijks leven. Zo bv. kunnen we wachtlijnfenomenen observeren op autowegen, in supermarkten, ... Wachtlijsten in ziekenhuizen is een ander voorbeeld. Algemeen kunnen we een *wachtlijnproces* definiëren als het wachten alvorens het verkrijgen van een zekere bediening.

Specifiek in telecommunicatienetwerken worden buffers gebruikt om informatie op te slaan die niet onmiddellijk naar haar bestemming kan verstuurd worden. Dit kan b.v. optreden doordat informatie van verschillende ingangslijnen gemultiplexeerd wordt op één uitgangslijn, informatie van een snelle ingangslijn moet getransporteerd worden naar een tragere uitgangslijn, ... Zonder buffers zou daardoor (te veel) informatie verloren gaan.

De entiteiten die aankomen in het wachtlijnsysteem, duiden we in het algemeen met de benaming *eenheden* aan doorheen dit proefschrift.

S.1.2 Belang van het wachtlijn- of buffergedrag

Het wachtlijngedrag is een belangrijk onderzoeksonderwerp. B.v., bij het ontwerpen van nieuwe wegen kan het voordelig zijn om vooraf uit te zoeken of een zeker ontwerp al dan niet aanleiding zal geven tot (grote) wachtlijnen (files) en/of - als er een aantal mogelijke ontwerpen zijn - welk ontwerp de kortste files veroorzaakt.

In het geval van een wachtlijst voor operaties kan het zijn dat sommige ingrepen zo vlug mogelijk moeten gebeuren wegens levensgevaar voor de patiënt. Het is dus belangrijk om de wachttijden te bestuderen zodat nodeloze sterfgevallen voorkomen kunnen worden.

In telecommunicatienetwerken tenslotte, is de manier waarop een buffer zich gedraagt cruciaal voor de prestatie van het netwerk, aangezien de prestatie en

de bedieningskwaliteit (*Engels*: Quality of Service of QoS) gerelateerd zijn aan dit buffergedrag. Informatie kan b.v. verloren gaan doordat een buffer vol is, of eenheden kunnen te veel vertragingstijd oplopen doordat ze te lang moeten wachten in de netwerkknooppunten alvorens verder gezonden te worden. De gevolgen voor de gebruikers zijn afhankelijk van de applicatie die zij gebruiken. Zo zijn lange vertragingstijden niet acceptabel voor reële-tijds-applicaties (*Engels*: real-time applications) - zoals telefonie - maar meestal wel aanvaardbaar voor data-applicaties (b.v. het versturen van bestanden). Omgekeerd zal verlies van informatie tot op zekere hoogte aanvaardbaar zijn voor telefonie terwijl dit onaanvaardbaar is voor de meeste data-applicaties.

S.1.3 Toevalsveranderlijken

Om het buffergedrag te bestuderen, definiëren we eerst een aantal ingangs- en uitgangsveranderlijken. De ingangsveranderlijken beschrijven de karakteristieken van het inkomende verkeer en worden bekend verondersteld doorheen dit proefschrift. De uitgangsveranderlijken beschrijven het buffergedrag en zijn dus te analyseren. Aangezien de verkeerskarakteristieken van een onzekere natuur zijn, definiëren we deze veranderlijken als *toevalsveranderlijken*. Elke veranderlijke wordt gekenmerkt door een probabiliteitsdistributie.

S.1.4 Analysetechnieken

Er zijn verschillende technieken om buffergedragingen te analyseren. Ze zijn ruwweg in te delen in vier verschillende categorieën: de analytische methode, de numerieke methode, simulaties en experimenten. Bij de eerste twee worden systeemvergelijkingen (uitgangsveranderlijken i.f.v. ingangsveranderlijken) opgesteld en deze worden respectievelijk analytisch en numeriek opgelost. Bij de laatste twee worden respectievelijk een computerversie en een ware versie van het desbetreffende (buffer)systeem geconstrueerd en worden de nodige resultaten bekomen door metingen uit te voeren.

Deze methoden hebben alle hun eigen voor- en nadelen. Het voordeel van de analytische methode is dat de afhankelijkheid van de prestatie van de verschillende ingangsparameters meestal onmiddellijk duidelijk is. Het grote nadeel van deze techniek is dat er meestal gebruik moet gemaakt worden van een vereenvoudigd wiskundig model zodanig dat het probleem analytisch oplosbaar is. De voor- en nadelen van de experimentele techniek zijn juist het tegenovergestelde, terwijl de numerieke techniek en de simulatie-aanpak - qua voor- en nadelen - tussen beide andere technieken in liggen.

In dit proefschrift zullen we gebruik maken van een analytische methode gebaseerd op *probabiliteitsgenererende functies* (*Engels:* probability generating functions of pgf's) om het desbetreffende wachtlijnmodel te analyseren.

S.1.5 Verschillende types verkeer

In vele wachtlijnstudies wordt aangenomen dat alle verkeer van hetzelfde type is. Het meeste verkeer is echter heterogeen van nature. Het verkeer kan dus opgesplitst worden in verschillende klassen, zowel naar karakteristieken als naar vereisten. Verkeerskarakteristieken kunnen inderdaad verschillend zijn (monotoon verkeer t.o.v. grillig verkeer, eenheden van constante lengte t.o.v. eenheden van variabele lengte, ...). Ten tweede kunnen ook de vereisten verschillen, bv., verschillende verliesvereisten, verschillende vertragingstijdvereisten, ... Een veelgebruikte classificatie in multimedianetwerken is dan ook reële-tijds-verkeer t.o.v. niet-reële-tijds-verkeer.

S.1.6 Schedulering van verschillende types verkeer

Als het verkeer (o.m.) geclassificeerd is op basis van vereisten, kunnen verschillende types van verkeer op verschillende manieren 'behandeld' worden in de buffersystemen. Zo b.v. kunnen de eenheden worden opgedeeld naargelang hun vertragingsgevoeligheid en kan verkeer van een vertragingsgevoelige klasse (gemiddeld gezien) voorrang verkrijgen op vertragingsongevoelig verkeer wat betreft het verzenden. In de Engelstalige tekst (sectie 1.6) worden een aantal van deze 'scheduleringsdisciplines' besproken. In dit proefschrift lichten we er één uit en analyseren deze grondig, nl., de *prioriteitsdiscipline* (*Engels:* priority discipline). Bij deze discipline wordt het verkeer verdeeld over een aantal klassen met verschillende prioriteit en eenheden van een bepaalde klasse kunnen alleen maar bediend worden als er geen eenheden van een klasse met hogere prioriteit aanwezig zijn.

S.1.7 Het wachtlijnmodel

We beschrijven de karakteristieken van het wachtlijnmodel dat we doorheen dit proefschrift gebruiken. De modellering kan opgesplitst worden in 3 delen, nl., de modellering van *buffersysteem zelf*, de modellering van het *aankomstproces* en de modellering van het *bedieningsproces*.

Het buffersysteem

We analyseren een *discrete-tijd* buffersysteem met *twee klassen*, *één bedieningsstation*, een *oneindige bufferlengte* en een *prioriteitsdiscipline*. Het bedieningsstation staat in voor de bediening van de eenheden. Eenheden die niet onmiddellijk kunnen verzonden worden, worden gebufferd. Aangezien we een discrete-tijd systeem analyseren, nemen we aan dat alle toevalsveranderlijken enkel discrete (niet-negatieve) waarden kunnen aannemen. De buffer is oneindig groot verondersteld, wat er op neerkomt dat er geen eenheden verloren kunnen gaan. Bediening kan enkel aanvangen op slotgrenzen. Daardoor

kan de bediening van een eenheid ten vroegste starten bij het begin van het slot na zijn aankomstslot. Het verkeer is verondersteld onderverdeeld te zijn in 2 klassen. De prioriteitsdiscipline zelf zal op het einde van deze subsectie in meer detail beschreven worden.

Het aankomstproces

Het aankomstproces wordt gekarakteriseerd door het aantal aankomsten per slot. Het aantal aankomsten van klasse-j gedurende het k-de slot wordt aangeduid met $a_{j,k}$, $j = 1, 2$. Doorheen het proefschrift wordt aangenomen dat de aantallen aankomsten van de verschillende klassen onafhankelijk en identisch gedistribueerd (*Engels*: independent and identically distributed of i.i.d.) zijn van slot-tot-slot en deze worden met de volgende gezamenlijke probabiliteitsgenererende functie gekarakteriseerd:

$$A(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{a_{1,k}} z_2^{a_{2,k}}\right]. \tag{S.1}$$

Belangrijk aan deze gezamenlijke pgf is dat ze incorporeert dat de aantallen aankomsten van de twee klassen in 1 slot stochastisch afhankelijk kunnen zijn. Dit soort aankomstprocessen wordt ook wel *gestructureerde input* (*Engels*: structured input) genoemd. Uit deze gezamenlijke pgf kunnen de marginale pgf's van het aantal aankomsten van klasse-1 en klasse-2 berekend worden:

$$\begin{aligned} A_j(z) &\triangleq \mathrm{E}[z^{a_{j,k}}] && \text{(S.2)} \\ &= A(z_1, z_2)|_{z_j = z, z_i = 1, i \neq j}, && \text{(S.3)} \end{aligned}$$

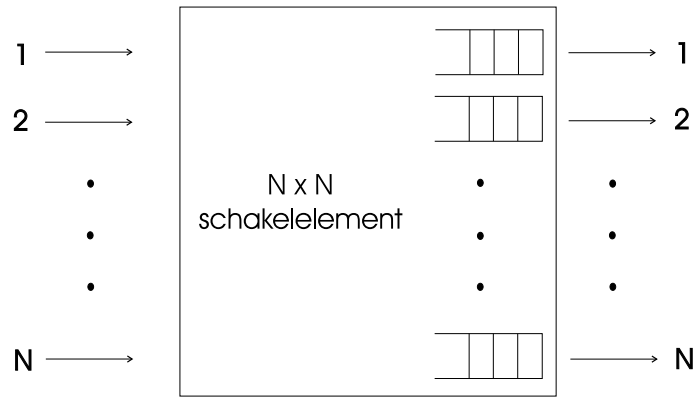
$j = 1, 2$, alsook de pgf van het totaal aantal aankomsten in een slot:

$$\begin{aligned} A_T(z) &\triangleq \mathrm{E}[z^{a_{1,k} + a_{2,k}}] && \text{(S.4)} \\ &= A(z, z). && \text{(S.5)} \end{aligned}$$

De (gemiddelde) aankomstintensiteit van klasse-j wordt genoteerd met λ_j en kan gevonden worden uit de corresponderende pgf als $A_j'(1)$. De totale aankomstintensiteit λ_T is dan de som van de klasse-1 en klasse-2 aankomstintensiteiten. In dit proefschrift zullen we - als voorbeeld - doorlopend een (tweedimensionaal) binomiaalverdeeld aantal per-slot aankomsten van beide klassen veronderstellen, i.e.,

$$A(z_1, z_2) = \left(1 + \sum_{j=1}^{2} \frac{\lambda_j}{N}(z_j - 1)\right)^N. \tag{S.6}$$

Dit is het aankomstproces dat waargenomen wordt aan de ingang van een willekeurige uitgangsbuffer in een $N \times N$ schakelelement met uitgangsbuf-



Figuur S.1: Een $N \times N$ schakelelement met uitgangsbuffers

fers (zie Figuur S.1), waarbij de aankomstprocessen aan de ingangen van het schakelelement gekenmerkt worden door (onafhankelijke en identische) Bernoulli-processen met parameter λ_T - een aankomende eenheid is van klasse-j met probabiliteit λ_j/λ_T, $j = 1, 2$ - en waarbij alle eenheden onafhankelijk en uniform naar de uitgangen geschakeld worden.

Het bedieningsproces

Het bedieningsproces heeft de volgende (algemeen gemaakte) beperkingen: de bedieningstijden zijn onderling onafhankelijk en bedieningstijden van eenheden van klasse-j zijn identisch gedistribueerd. Noteren we de bedieningstijd van een willekeurige klasse-j eenheid als s_j, dan is de pgf van de klasse-j bedieningstijden gegeven door

$$S_j(z) \triangleq \mathrm{E}[z^{s_j}]. \tag{S.7}$$

De gemiddelde bedieningstijd van een klasse-j eenheid - genoteerd als μ_j - is gegeven door $S_j'(1)$. De aankomstbelasting van klasse-j is dus gegeven door $\rho_j = \lambda_j \mu_j$ en de totale aankomstbelasting ρ_T is de som van de klasse-1 en klasse-2 belasting.

De prioriteitsdiscipline

Tenslotte bespreken we de prioriteitsdiscipline. Er wordt aangenomen dat klasse-1 eenheden voorrang hebben op de klasse-2 eenheden bij de bediening. Binnen één klasse nemen we aan dat de discipline FCFS (*Engels*: First Come First Served) is, wat betekent dat de eenheden van een specifieke klasse bediend worden in volgorde van aankomst. Dus als het bedieningsstation

vrijkomt, wordt een klasse-1 eenheid bediend en enkel als er geen klasse-1 eenheden zijn, kan de bediening van een klasse-2 eenheid aanvangen. Er kunnen twee soorten prioriteitsdisciplines onderscheiden worden, nl., de *niet-preëmptieve prioriteitsdiscipline* en de *preëmptieve prioriteitsdiscipline*. Bij de niet-preëmptieve (*Engels*: non-preemptive of NP) prioriteitsdiscipline worden bedieningen nooit onderbroken. Dus als er klasse-1 eenheden aankomen gedurende een klasse-2 bediening, kan deze bediening niet onderbroken worden en moeten de klasse-1 eenheden (ten minste) wachten tot het einde van die bediening. Bij de preëmptieve prioriteitsdisciplines daarentegen, wordt een klasse-2 bedieningstijd wel onderbroken door nieuwe klasse-1 aankomsten. Bij deze laatste kunnen nog drie soorten onderscheiden worden, nl., de *preëmptieve met voortzetting* (*Engels*: preemptive resume of PR), de *preëmptieve met verschillende herhaling* (*Engels*: preemptive repeat different of PRD) en de *preëmptieve met identieke herhaling* (*Engels*: preemptive repeat identical of PRI) prioriteitsdisciplines. Bij de eerste soort wordt de onderbroken bediening voortgezet na de onderbreking, of m.a.w., enkel het nog niet bediende gedeelte van de eenheid moet na de onderbreking nog bediend worden. In de twee preëmptieve disciplines met herhaling moet de *volledige* eenheid opnieuw bediend worden. Bij deze met verschillende herhaling kan de bedieningstijd veranderen na een onderbreking - er wordt dan een nieuw monster van de klasse-2 bedieningstijden genomen - terwijl bij deze met identieke herhaling de bedieningstijden identiek blijven na een onderbreking. Dus een eenheid heeft telkens dezelfde bedieningstijd bij elke bedieningspoging.

Merk tenslotte op, dat als de bedieningstijden van klasse-2 deterministisch gelijk zijn aan 1 slot, alle besproken prioriteitsdisciplines identiek zijn. Dit is wegens het feit dat een klasse-2 eenheid het systeem sowieso verlaat op het einde van zijn bedieningsslot, of er nu klasse-1 eenheden aankomen of niet gedurende dat slot.

S.1.8 Typische resultaten

Toevalsveranderlijken

Voor de verschillende wachtlijnmodellen bestudeerd in dit proefschrift, worden telkens een aantal toevalsveranderlijken gedefinieerd en geanalyseerd. Deze toevalsveranderlijken worden allen in regime geanalyseerd in dit proefschrift. Een eerste is de *systeembezetting* bij het begin van een willekeurig slot. Dit is het aantal eenheden dat aanwezig is in het systeem bij het begin van dit slot. Hierbij worden zowel de systeembezettingen van beide klassen afzonderlijk - genoteerd met u_1 en u_2 - gedefinieerd als de totale systeembezetting $u_T = u_1 + u_2$.

Gerelateerde toevalsveranderlijken zijn de *wachtlijnbezettingen* bij het begin van een willekeurig slot, i.e., het aantal eenheden in de wachtlijn bij het begin van een willekeurig slot. Het verschil tussen de systeembezettingen en de

wachtlijnbezettingen is dat bij de eerste de eenheid die in bediening is wordt meegerekend, terwijl dit niet het geval is bij de laatste. Ook de wachtlijnbezetting kan per klasse gedefinieerd worden - genoteerd met q_1 en q_2 - alsook voor de totale wachtlijn - genoteerd met q_T.

Een derde soort veranderlijke is het *onvoltooid werk*. Het totale onvoltooid werk bij het begin van een slot w_T is gedefinieerd als het aantal slots nodig om de bediening van alle eenheden in het systeem bij het begin van dat slot af te werken, waarbij men aanneemt dat er geen nieuwe eenheden meer zouden aankomen. Het onvoltooid werk van klasse-j, w_j, ($j = 1, 2$) is dan het aantal slots van dit totale onvoltooid werk dat gespendeerd wordt aan bedieningen van klasse-j eenheden.

Tenslotte definiëren we nog de *vertragingstijd* en de *wachttijd*. De vertragingstijd van een eenheid wordt gedefinieerd als het aantal slots dat deze zich in het buffersysteem bevindt, terwijl de wachttijd het aantal slots is dat de eenheid zich in de wachtlijn bevindt. De bedieningstijd maakt dus deel uit van de vertragingstijd terwijl deze geen deel uitmaakt van de wachttijd. We definiëren dus zowel de klasse-j vertragings- en wachttijd - respectievelijk genoteerd als d_j en t_j - alsook de vertragings- en de wachttijd van een willekeurige (klasse-1 of klasse-2) eenheid - respectievelijk d en t.

In deze Nederlandstalige samenvatting zullen we enkel ingaan op de analyse van de systeembezettingen en vertragingstijden, maar in het Engelstalig deel is er ook op de analyse van de andere gedefinieerde toevalsveranderlijken (meestal kort) ingegaan.

Probabiliteitsgenererende functies (pgf's)

In dit proefschrift maken we veelvuldig gebruik van probabiliteitsgenererende functies. We zullen - voor de verschillende modellen - telkens de tweedimensionale pgf's berekenen van de klasse-1 en klasse-2 systeembezettingen bij het begin van een willekeurig slot in regime. Uit deze tweedimensionale pgf's kunnen dan de marginale pgf's van de toevalsveranderlijken berekend worden, alsook de pgf van de som van beiden (of dus de pgf's van de totale systeembezetting, totale wachtlijnbezetting en totaal onvoltooid werk respectievelijk). Verder zullen we de pgf's van de vertragingstijden en wachttijden van klasse-1, klasse-2 en willekeurig geselecteerde eenheden berekenen.

Momenten

Uit de verkregen pgf's kunnen we uiteindelijk de effectieve prestatiematen berekenen, te beginnen met de *momenten* van de verschillende toevalsveranderlijken. De gemiddelde waarde en variantie van een veranderlijke X met pgf $X(z)$ zijn respectievelijk gelijk aan

$$\mathrm{E}[X] = X'(1) \tag{S.8}$$

$$\text{Var}[X] = X''(1) + X'(1) - (X'(1))^2. \tag{S.9}$$

Dus door het afleiden van de genererende functies en het evalueren in $z = 1$ kunnen alle (centrale) momenten van de toevalsveranderlijken - waarvan de genererende functies berekend zijn - gevonden worden. Verder kunnen van toevalsveranderlijken waarvan de gezamenlijke genererende functie gevonden is de kruismomenten berekend worden. Zo worden b.v. de covariantie en de correlatiecoëfficiënt van twee toevalsveranderlijken X_1 en X_2, met gezamenlijke genererende functie $X(z_1, z_2)$, respectievelijk gegeven door

$$\text{Cov}[X_1, X_2] \triangleq \text{E}\left[(X_1 - \text{E}[X_1])(X_2 - \text{E}[X_2])\right] \tag{S.10}$$

$$= \left.\frac{\partial^2 X(z_1, z_2)}{\partial z_1 \partial z_2}\right|_{z_1=z_2=1} - X_1'(1)X_2'(1) \tag{S.11}$$

$$\text{Corr}[X_1, X_2] \triangleq \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var}[X_1]\text{Var}[X_2]}}. \tag{S.12}$$

Massafuncties en staartprobabiliteiten

Een nog belangrijkere prestatiemaat is de probabiliteitsmassafunctie (*Engels*: probability mass function of pmf) van een toevalsveranderlijke X:

$$x(n) = \text{Prob}[X = n], \tag{S.13}$$

voor $n = 0, 1, \ldots$. Merk op dat de pgf $X(z)$ van X niets anders is dan de z-getransformeerde van de pmf $x(n)$, i.e.,

$$X(z) = \sum_{n=0}^{\infty} x(n) z^n. \tag{S.14}$$

Omgekeerd worden deze $x(n)$ uit $X(z)$ gevonden als:

$$x(n) = \left.\frac{1}{n!}\frac{d^n X(z)}{dz^n}\right|_{z=0}. \tag{S.15}$$

Dus eens $X(z)$ berekend is, kan d.m.v. deze laatste formule $x(n)$ (voor alle n) in principe berekend worden. Aangezien we echter meestal geïnteresseerd zijn in de $x(n)$ voor grote n, is dit geen praktisch werkbare formule, wegens het feit dat $X(z)$ n maal moet afgeleid worden voor de berekening van $x(n)$. Daarom worden in de bestaande literatuur meestal benaderende technieken gebruikt die exact zijn voor $n \to \infty$ (de zogenaamde *staartprobabiliteiten*). In dit proefschrift zullen we veelvuldig gebruik maken van enkele theorema's, b.v., het theorema van Darboux (theorema 1.1 in de Engelstalige tekst), die de staartprobabiliteiten geven als (een deel van) het gedrag van de pgf's in hun

dominante singulariteit(en) (dit zijn de singulariteiten met kleinste norm) gekend is. Merk op dat 1 van die dominante singulariteiten op de positieve reële as ligt. In de literatuur wordt meestal verondersteld dat dit de enige dominante singulariteit is. We hebben doorheen dit proefschrift deze veronderstelling ook gemaakt.

In dit proefschrift worden de staartprobabiliteiten van de relevante toevalsveranderlijken berekend voor aankomstprocessen en bedieningsprocessen die aan bepaalde voorwaarden voldoen (die er meestal op neerkomen dat de (marginale) pgf's en hun afgeleiden naar oneindig moeten gaan voor z gelijk aan hun convergentiestraal). Deze voorwaarden zijn voldaan voor de meeste 'normale' aankomst- en bedieningsprocessen, maar indien dit niet zo is, zijn de gebruikte technieken meestal uitbreidbaar.

S.1.9 Overzicht van dit proefschrift

We beëindigen deze inleidende sectie met een kort overzicht (van het Nederlandstalig deel) van dit proefschrift. In de volgende sectie bespreken we het geval van vaste bedieningstijden van 1 slot voor alle eenheden. Aangezien de klasse-2 eenheden vaste bedieningstijden hebben van 1 slot, hoeven we in dit geval dus geen opsplitsing te doen van de prioriteitsdisciplines. We zullen de invloed van de aankomstkarakteristieken op de prestatiematen nagaan via enkele numerieke voorbeelden. In sectie S.3 bespreken we hoe de niet-preëmptieve prioriteitsbuffer, de preëmptieve prioriteitsbuffer met voortzetting en de preëmptieve prioriteitsbuffer met verschillende en identieke herhaling geanalyseerd kunnen worden, als de bedieningstijden van de pakketten algemeen gedistribueerd zijn (en eventueel verschillend voor beide prioriteitsklassen). We tonen d.m.v. een aantal numerieke voorbeelden ook de invloeden van en de verschillen tussen de verschillende disciplines aan. Tenslotte besluiten we deze Nederlandstalige sectie en belichten kort enkele mogelijke uitbreidingen.

S.2 Deterministische bedieningstijden van 1 slot

In eerste instantie hebben we een model bestudeerd waarbij de bedieningstijden van beide klassen deterministisch gelijk zijn aan één slot. Dit is een model dat veelvuldig gebruikt wordt in discrete-tijd wachtlijnmodellen, aangezien het enerzijds het eenvoudigste model is voor de bedieningstijden en aangezien het anderzijds in de praktijk vaak voorkomt dat alle eenheden dezelfde bedieningstijd hebben. In een telecommunicatiecontext b.v., wordt dit soort model gebruikt om wachtlijngedragingen in ATM (Asynchronous Transfer Mode) netwerken te bestuderen. In ATM netwerken zijn de *cellen* - in ATM worden de eenheden aangeduid door cellen - allen van dezelfde lengte en de tijd nodig om een cel te verzenden wordt als slotlengte genomen.

Discrete-tijd prioriteitssystemen met bedieningstijden van één slot en geen correlatie tussen de aankomstprocessen van de verschillende prioriteitsklassen zijn bestudeerd door Hashida and Takahashi [1991], Schormans et al. [1991], Takine et al. [1994b], Choi et al. [1998a], Shakkottai and Srikant [2001], Xabier Albizuri et al. [2003] en Mehmet Ali and Song [2004], waarbij de verschillende artikels zowel verschillen in de modellering van het aankomstproces als in de gebruikte oplossingstechnieken en behaalde resultaten.

Discrete-tijd prioriteitssystemen met bedieningstijden van één slot en correlatie tussen de aankomstprocessen van de verschillende prioriteitsklassen zijn geanalyseerd in [Sidi and Segall 1983, Chang and Harn 1992, Khamisy and Sidi 1992, Laevens and Bruneel 1998, Walraevens and Bruneel 1999] en [Walraevens et al. 2003c], waarbij in al deze artikels de pgf-methode gebruikt is, maar waarbij er onderling verschillen zijn in het aantal bedieningsstations en in de behaalde resultaten.

Alhoewel dit initieel bestudeerde model - met bedieningstijden gelijk aan één slot - een speciaal geval is van het model bestudeerd in [Laevens and Bruneel 1998] - waarin een prioriteitsbuffer met meerdere bedieningsstations geanalyseerd is - en van de verdere modellen bestudeerd in dit proefschrift - waarbij meer algemene bedieningstijden aangenomen worden - vinden we het nuttig om de analyse van dit model in dit proefschrift uitvoerig te beschrijven. Dit geeft ons de kans om de technieken die we doorheen het proefschrift gebruiken vooreerst uit te leggen voor een vrij eenvoudig model. Verder geeft dit eenvoudig model ons ook de kans om de (pure) invloed van het aankomstproces op de prestatiematen te bestuderen, aangezien het aankomstproces de enige 'stochastische input' is in dit model. De analyse van dit model is het onderwerp van het tweede hoofdstuk in het Engelstalige gedeelte.

S.2.1 Berekening pgf's

Systeembezettingen

Allereerst wordt de gezamenlijke pgf $U(z_1, z_2)$ van de systeembezettingen van klasse-1 en klasse-2 bij het begin van een willekeurig slot in regime berekend, door uit te gaan van de systeemvergelijkingen (systeembezetting bij begin van een slot i.f.v. de systeembezetting bij het begin van het vorige slot en de nieuwe aankomsten gedurende dit laatste slot). Dit werkt aangezien de bedieningstijden gelijk zijn aan één slot. Daardoor vormen de systeembezettingen van klasse-1 en klasse-2 bij het begin van opeenvolgende slots een Markov keten of m.a.w. deze toevalsveranderlijken bij het begin van een slot zijn gekend als we diezelfde toevalsveranderlijken kennen bij het begin van het vorige slot - alsmede de ingangsveranderlijken (in dit geval zijn dit het aantal klasse-1 en klasse-2 aankomsten gedurende het vorige slot). $U(z_1, z_2)$ wordt dan gevonden, door onder meer gebruik te maken van Rouché's theorema (zie Appen-

dix) en de normalisatievoorwaarde ($U(1,1) = 1$). We verkrijgen uiteindelijk

$$U(z_1, z_2) = (1 - \lambda_T) \frac{A(z_1, z_2)(z_1 - Y(z_2))(z_2 - 1)}{(z_1 - A(z_1, z_2))(z_2 - Y(z_2))}, \tag{S.16}$$

waarbij

$$Y(z) \triangleq A(Y(z), z). \tag{S.17}$$

Belangrijk hierbij is dat $Y(z)$ enkel impliciet gedefinieerd is (voor algemene aankomstprocessen). Daardoor is $U(z_1, z_2)$ niet expliciet berekend, maar we tonen aan dat dit geen probleem vormt om de prestatiematen te berekenen (zie verder). De gezamenlijke pgf $U(z_1, z_2)$ is vervolgens het uitgangspunt van alle verdere berekeningen. Zo b.v. zijn hieruit de marginale pgf's van de klasse-1, klasse-2 en totale bufferbezettingen eenvoudig te vinden.

Vertragingstijden

Verder worden de pgf's van de vertragingstijden van klasse-1 en klasse-2 cellen (afzonderlijk) berekend. Dit gebeurt door de vertragingstijden van een willekeurige klasse-1 cel of willekeurige klasse-2 cel uit te drukken i.f.v. de systeembezettingen bij het begin van hun aankomstslots en deze te z-transformeren. Bij de klasse-2 vertragingstijd voeren we dan de notie van de *fundamentele periode* (*Engels*: sub-busy period) in. Deze is ruwweg gedefinieerd als het aantal slots dat nodig is om het aantal wachtende cellen die bediend moeten worden vóór een bepaalde klasse-2 cel met één te verminderen. Merk op dat deze fundamentele periode niet noodzakelijk gelijk is aan 1 aangezien klasse-1 cellen die aankomen terwijl de klasse-2 cel zich in de wachtlijn bevindt *voor* deze cel bediend worden. Uiteindelijk verkrijgen we de volgende pgf's voor de klasse-1 en klasse-2 vertragingstijden:

$$D_1(z) = \frac{1 - \lambda_1}{\lambda_1} \frac{z(A_1(z) - 1)}{z - A_1(z)} \tag{S.18}$$

$$D_2(z) = \frac{1 - \lambda_T}{\lambda_2} \frac{zA_T(V(z)) - V(z)}{V(z) - A_T(V(z))}, \tag{S.19}$$

met

$$V(z) = zA_1(V(z)), \tag{S.20}$$

de pgf van de fundamentele periodes. Merk op dat $V(z)$ - net als $Y(z)$ - enkel impliciet gedefinieerd is.

S.2.2 Berekening prestatiematen

Zoals in de inleiding besproken kunnen uit de behaalde pgf's de momenten en staartprobabiliteiten van de verschillende toevalsveranderlijken berekend worden. Het enige overblijvende probleem is het feit dat impliciet gedefinieerde functies voorkomen in de pgf's van de klasse-2 toevalsveranderlijken. We bespreken kort het effect hiervan op de berekening van momenten en staartprobabiliteiten.

Momenten

Bij de berekening van de momenten moeten de genererende functies een aantal maal afgeleid worden en vervolgens geëvalueerd worden in $z = 1$. Aangezien $Y(z)$ en $V(z)$ pgf's (blijken te) zijn, zijn deze gelijk aan 1 in $z = 1$. Doordat we de expliciete waarde van deze functies dus weten voor $z = 1$, kunnen we ook alle afgeleiden van deze functies, geëvalueerd in 1, expliciet berekenen. We besluiten dus dat de berekeningen van de momenten uit de pgf's geen probleem vormen.

Staartprobabiliteiten

Wat de berekening van de staartprobabiliteiten betreft, gaan we eerst in op de staartprobabiliteiten van de toevalsveranderlijken waarbij zich geen impliciet gedefinieerde functies bevinden (vooral de "klasse-1 toevalsveranderlijken"). Noteren we de pgf van een toevalsveranderlijke met $X(z)$. Deze heeft een enkelvoudige dominante pool z_* op de positieve reële as (> 1). Deze pgf is dus in de buurt van deze pool benaderend te schrijven als

$$X(z) \approx \frac{K}{z_* - z}, \tag{S.21}$$

waarbij K berekend wordt door de limiet van $X(z)$ te berekenen voor $z \to z_*$. Doordat we het benaderende gedrag van de pgf in zijn dominante pool kennen, kunnen we de staartprobabiliteiten berekenen - gebruik makende van Darboux's theorema - en we verkrijgen het volgende *geometrische staartgedrag*

$$x(n) \approx K z_*^{-n-1}, \tag{S.22}$$

met $x(n)$ de probabiliteitsdistributie horende bij $X(z)$. Deze benadering wordt beter naarmate n groter is en is exact voor $n \to \infty$.

De staartprobabiliteiten berekenen van een toevalsveranderlijke wiens pgf impliciet gedefinieerd is, $Y(z)$ b.v., is gecompliceerder. Dit is vooral omdat deze functies doorgaans dominante singulariteiten hebben die geen geïsoleerde polen zijn. In hun dominante singularteit blijven deze functies zelf eindig, terwijl

hun afgeleiden er oneindig worden. Dit punt is een vertakkingspunt. Gebruiken we $Y(z)$ hier verder als voorbeeld, dan kunnen we deze pgf in de buurt van haar dominante singulariteit z_B schrijven als

$$Y(z) \approx Y(z_B) - K_Y \left(z_B - z\right)^{1/2}, \tag{S.23}$$

waarbij K_Y gevonden wordt door $Y(z)$ en zijn afgeleiden te evalueren in z_B. Opnieuw gebuik makend van Darboux's theorema vinden we het volgende *niet-geometrische gedrag* voor de distributie behorende bij $Y(z)$:

$$y(n) \approx \frac{K_Y}{2}\sqrt{\frac{z_B}{\pi}} n^{-3/2} z_B^{-n}. \tag{S.24}$$

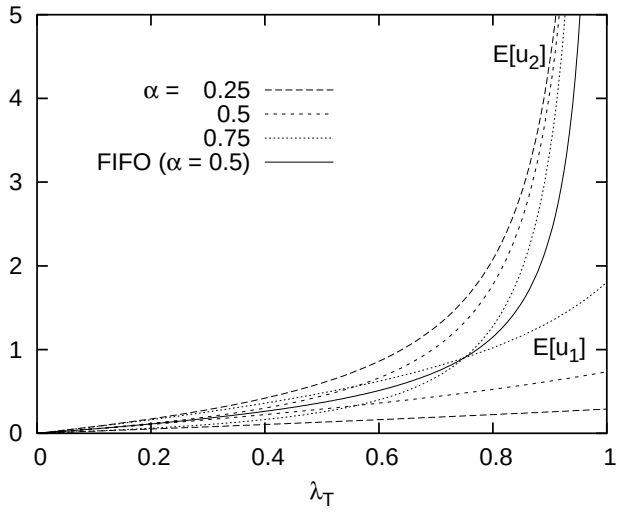
Uiteindelijk kunnen de staartprobabiliteiten van de toevalsveranderlijken berekend worden wiens pgf's een impliciet gedefinieerde functie - b.v. $Y(z)$ - bevatten. Noteren we zo een pgf opnieuw met $X(z)$. Deze heeft dan twee singulariteiten die een rol spelen, nl. een enkelvoudige pool - opnieuw genoteerd met z_* - en het vertakkingspunt z_B van de impliciet gedefinieerde functie. Drie gevallen kunnen zich dan onderscheiden, nl., z_* is dominant, $z_* = z_B$ is dominant en z_B is dominant. In deze drie gevallen kan de genererende functie $X(z)$ in de buurt van zijn dominante singulariteit geschreven worden als

$$X(z) \approx \begin{cases} \dfrac{K^{(1)}}{z_* - z} & \text{als } z_* \text{ dominant} \\ \dfrac{K^{(2)}}{(z_B - z)^{1/2}} & \text{als } z_* = z_B \text{ dominant} \\ X(z_B) - K^{(3)}(z_B - z)^{1/2} & \text{als } z_B \text{ dominant,} \end{cases} \tag{S.25}$$

waarbij de $K^{(i)}$ $(i = 1, 2, 3)$ kunnen berekend worden door de limiet van $X(z)$ voor z gaande naar zijn dominante singulariteit te bepalen. Door Darboux's theorema te gebruiken worden de staartprobabiliteiten gevonden:

$$x(n) \approx \begin{cases} K^{(1)} z_*^{-n-1} & \text{als } z_* \text{ dominant} \\ \dfrac{K^{(2)} n^{-1/2} z_B^{-n}}{\sqrt{\pi z_B}} & \text{als } z_* = z_B \text{ dominant} \\ \dfrac{K^{(3)} n^{-3/2} z_B^{-n}}{2\sqrt{\pi / z_B}} & \text{als } z_B \text{ dominant.} \end{cases} \tag{S.26}$$

We besluiten dus dat de invloed van een impliciet gedefinieerde functie in de uitdrukking van een pgf is dat het staartgedrag niet langer noodzakelijkerwijs geometrisch is.



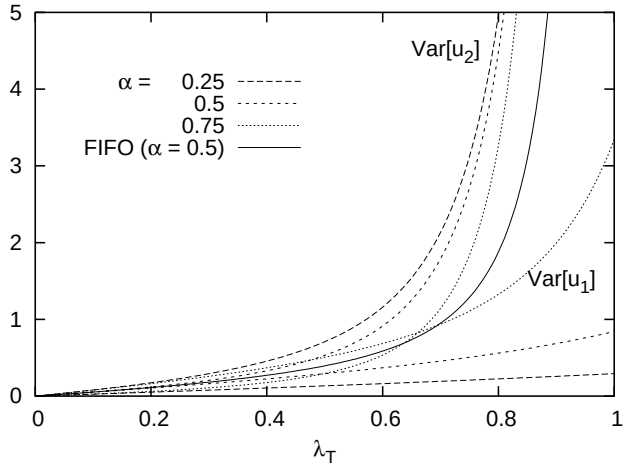
Figuur S.2: Gemiddelde systeembezettingen versus de totale aankomstintensiteit

S.2.3 Numerieke voorbeelden

In deze paragraaf zullen we kort de invloed van enkele parameters op de prestatiematen tonen. We zullen dit doen a.d.h.v. het schakelelement met uitgangsbuffers zoals eerder vermeld (zie Figuur S.1). De gezamenlijke pgf van het aantal klasse-j aankomsten ($j = 1, 2$) aan een willekeurige uitgangsbuffer is gegeven door (S.6). We definiëren α als de fractie klasse-1 belasting van de totale belasting (in dit model is dit gelijk aan λ_1/λ_T). Zonder andere vermelding veronderstellen we N - het aantal ingangen van het schakelelement - steeds gelijk aan 16.

Figuren S.2 en S.3 tonen de gemiddelde waarden en varianties van de systeembezettingen van klasse-1 en klasse-2 als functies van de totale aankomstintensiteit λ_T voor verschillende waarden van α. We hebben deze grootheden ook uitgezet voor een FIFO discipline i.p.v. een prioriteitsdiscipline voor $\alpha = 0.5$. Merk op dat de gemiddelde waarden en varianties van de bufferbezettingen van beide klassen in dit geval gelijk zijn. Uit deze figuren kan duidelijk de invloed gezien worden van de prioriteitsdiscipline: de gemiddelde waarde en variantie van de klasse-1 systeembezetting worden gereduceerd door de prioriteitsdiscipline, terwijl het tegengestelde geldt voor de klasse-2 cellen, en dit vooral bij een hoge aankomstintensiteit.

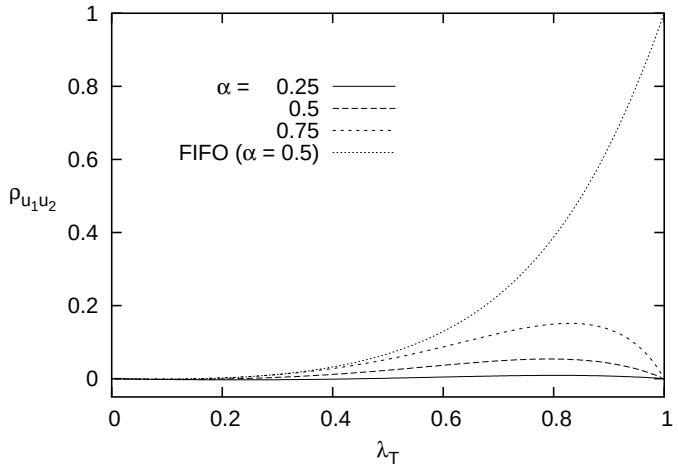
Figuur S.4 toont de correlatiecoëfficiënt $\rho_{u_1u_2}$ van de klasse-1 en klasse-2 systeembezettingen bij het begin van een slot als een functie van λ_T voor verschillende waarden van α (alsmede diezelfde correlatiecoëfficiënt als de discipline FIFO is en $\alpha = 0.5$). $\rho_{u_1u_2}$ is licht negatief voor kleine λ_T, maar wordt positief



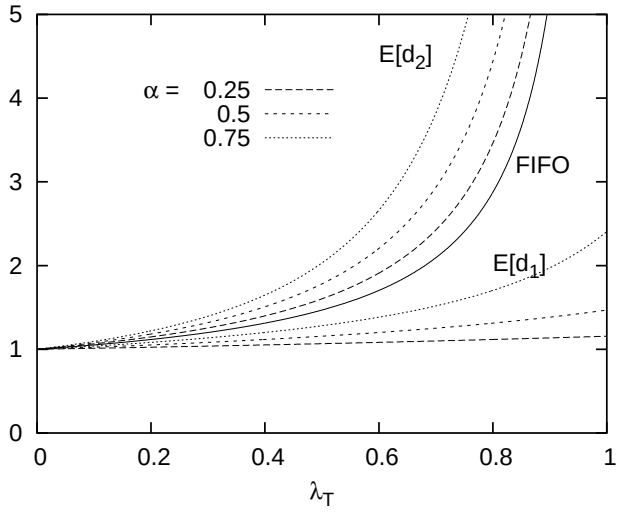
Figuur S.3: Varianties van de systeembezettingen versus de totale aankomstintensiteit

voor grotere λ_T. De reden daarvoor is dat er twee tegenwerkende effecten optreden. Het eerste is de negatieve correlatie tussen het aantal aankomsten van klasse-1 en klasse-2 in een slot (aangezien het aantal ingangen van het schakelelement eindig is) die doorgegeven wordt aan de correlatie tussen de systeembezettingen. Dit effect is vooral belangrijk bij lage aankomstintensiteiten aangezien er dan zo goed als geen buffering optreedt. De correlatie tussen de bufferbezettingen is dan zo goed als het directe gevolg van de correlatie tussen het aantal klasse-1 en klasse-2 aankomsten binnen een slot. De tweede invloed is de prioriteitsdiscipline zelf: als λ_1 en λ_T toenemen, komen er steeds meer cellen aan en de aanwezigheid van klasse-1 cellen verhindert de bediening van klasse-2 cellen, waardoor de correlatie positief wordt: hoe groter de klasse-1 systeembezetting, hoe groter de klasse-2 systeembezetting. Merk op dat als $\lambda_T \to 1$ de buffer van klasse-2 onstabiel wordt en de systeembezetting van klasse-2 naar oneindig gaat (onafhankelijk van de klasse-1 systeembezetting). Daardoor gaat $\rho_{u_1 u_2}$ naar 0.

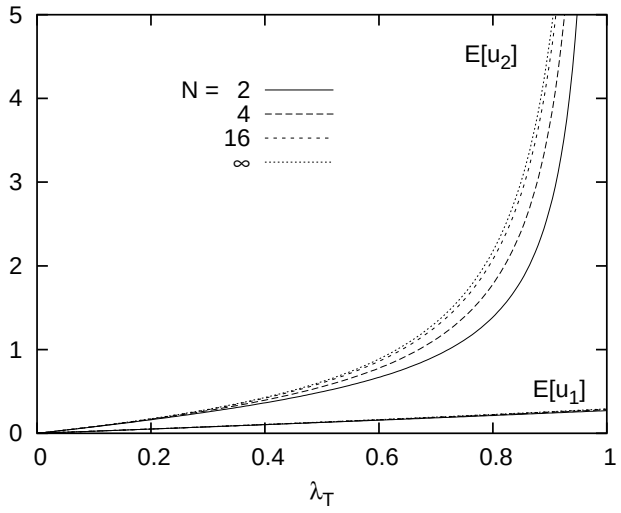
Figuur S.5 toont de gemiddelde klasse-1 en klasse-2 vertragingstijden als functies van λ_T voor verschillende waarden van α. Opnieuw werd ook de vertragingstijd getoond in het geval van een FIFO discipline (die aanleiding geeft tot gelijk gedistribueerde vertragingstijden voor klasse-1 en klasse-2 in dit specifiek geval). Het is duidelijk dat de gemiddelde klasse-1 vertragingstijd gereduceerd wordt, terwijl de gemiddelde klasse-2 vertragingstijd stijgt als gevolg van de prioriteitsdiscipline. Dit is natuurlijk precies waarom een prioriteitsdiscipline wordt ingevoerd. Tenslotte kunnen we ook opmerken dat zowel de gemiddelde klasse-1 als klasse-2 vertragingstijden stijgen met stijgende α, of m.a.w., dat de prioriteitsdiscipline het best werkt als het aantal cellen die prioriteit krijgen zo klein mogelijk gehouden wordt.



Figuur S.4: Correlatiecoëfficiënt van de klasse-1 en klasse-2 systeembezettingen versus de totale aankomstintensiteit



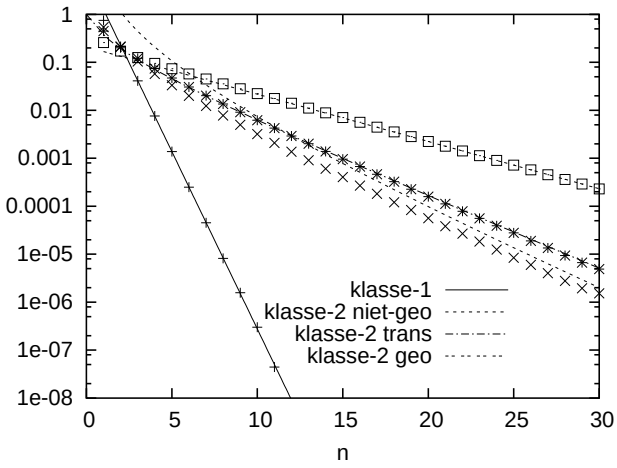
Figuur S.5: Gemiddelde vertragingstijden versus de totale aankomstintensiteit



Figuur S.6: Invloed van de tweede-orde momenten van het aankomstproces op de gemiddelde systeembezettingen

In Figuur S.6 worden de gemiddelde systeembezettingen van klasse-1 en klasse-2 getoond als functies van λ_T voor $\alpha = 0.25$ en voor verschillende waarden van N (het aantal ingangen van het bestudeerde schakelelement). Door N te variëren worden de tweede-orde karakteristieken - varianties en covarianties - van de klasse-1 en klasse-2 aankomsten in een slot gevarieerd. Wanneer N stijgt, neemt de variantie van de klasse-1 en klasse-2 aankomsten toe en wordt de covariantie minder negatief. Daardoor stijgen ook de gemiddelde systeembezettingen van klasse-1 en klasse-2 (zie figuur) en deze toename is vooral belangrijk in het geval van de gemiddelde klasse-2 systeembezetting en wanneer N klein is (de stijging als het aantal schakelelementingangen van 2 naar 4 gaat is b.v. een stuk groter dan wanneer deze van 16 naar ∞ gaat).

Tenslotte toont Figuur S.7 de staartprobabiliteiten van de klasse-1 en klasse-2 vertragingstijden voor $\lambda_1 = 0.4$ en $\lambda_2 = 0.1, 0.21$ en 0.4. De λ_2 zijn zodanig gekozen dat de drie gedragingen van de staartprobabiliteiten voor de klasse-2 vertragingstijden gerepresenteerd zijn, nl. respectievelijk niet-geometrisch gedrag, het transitiegedrag en geometrisch gedrag. Aangezien de staartprobabiliteiten benaderingen zijn (voor eindige n) hebben we deze gevalideerd door de distributies te berekenen uit simulaties (aangegeven door de markeringen in de figuur). Deze tonen aan dat de benaderende waarden heel dicht tegen de exacte waarden liggen.



Figuur S.7: Staartprobabiliteiten van de klasse-1 en klasse-2 vertragingstijden voor een aantal combinaties van klasse-1 en klasse-2 aankomstintensiteiten

S.3 Algemene bedieningstijden

In het vervolg van dit proefschrift veronderstellen we dat de klasse-1 en klasse-2 bedieningstijden willekeurige distributies hebben. Deze kunnen tevens verschillend zijn, i.e., de distributie van de bedieningstijden van klasse-1 eenheden kan verschillend zijn van deze van de klasse-2 eenheden.

In hedendaagse multimedia pakketgebaseerde telecommunicatienetwerken b.v., hebben de eenheden doorgaans niet allemaal dezelfde lengte. Het model met constante bedieningstijden van één slot is dan ook te restrictief om de prestatiematen in deze netwerken te analyzeren. Prioriteitsdisciplines in deze multimedianetwerken zijn een belangrijk onderzoeksonderwerp. In [Xiao and Ni 1999] wordt bv. een prioriteitsdiscipline voorgesteld in het gedifferentieerde bedieningsmodel (*Engels:* differentiated service model). Hierbij wordt verkeer van één klasse (de Premium Bedieningsklasse) een hogere prioriteit gegeven over het andere verkeer. We zullen in deze sectie de pakketgebaseerde terminologie overnemen en de eenheden 'pakketten' noemen.

Zoals in subsectie S.1.7 vermeld is, kunnen verschillende soorten prioriteitsdisciplines gedefinieerd worden wanneer de klasse-2 bedieningstijden meer dan één slot kunnen bedragen, nl. de niet-preëmptieve (NP) prioriteitsdiscipline, de preëmptieve prioriteitsdiscipline met voortzetting (PR), de preëmptieve prioriteitsdiscipline met verschillende herhaling (PRD) en de preëmptieve prioriteitsdiscipline met identieke herhaling (PRI). Bij de NP prioriteitsdisciplines kan de bediening van eenheden niet onderbroken worden, terwijl bij de PR, PRD en PRI prioriteitsdisciplines nieuw aankomende klasse-1 eenheden een aan de gang zijnde klasse-2 bediening wel onderbreken. Bij

de PR prioriteitsdiscipline wordt de onderbroken bediening van de klasse-2 eenheid voortgezet, terwijl deze herhaald wordt bij de PRD en PRI disciplines. Het verschil tussen deze laatste twee is dat de herhaalde bediening een nieuwe monsterwaarde aanneemt bij de PRD discipline terwijl deze identiek blijft bij de PRI prioriteitsdiscipline. Deze vier prioriteitsdisciplines zijn elk op hun beurt geanalyzeerd in dit proefschrift. In het Engelstalig gedeelte van dit proefschrift zijn de analyses en berekeningen van de prestatiematen te vinden. In hoofdstuk 3 hebben we de analyse van de NP prioriteitswachtlijn beschreven, in hoofdstuk 4 de analyse van de PR prioriteitswachtlijn en in hoofdstuk 5 de analyses van de PRD en PRI prioriteitswachtlijnen.

Discrete-tijd wachtlijnen met een prioriteitsdiscipline en zonder correlatie in de aankomstprocessen van de verschillende prioriteitsklassen zijn bestudeerd door o.a. Rubin and Tsai [1989], Chen and Guérin [1991], Mukherjee et al. [1995], Choi et al. [1997], Lee et al. [1998], Wang et al. [2000], Lee [2001], Lee et al. [2003], Fiems et al. [2004] en Fiems [2004]. In [Rubin and Tsai 1989] zijn de (gemiddelde) wacht- en vertragingstijden geanalyzeerd in NP en PR prioriteitswachtlijnen gebruik makend van pgf's. ATM schakelelementen met buffers aan de ingang en een PR prioriteitsdiscipline zijn bestudeerd in [Chen and Guérin 1991] en [Lee et al. 1998]. In [Choi et al. 1997] is dit ook bestudeerd alsmede het geval dat een NP prioriteitsdiscipline gebruikt wordt i.p.v. een PR prioriteitsdiscipline. In [Mukherjee et al. 1995] wordt een PRI (of PRD) prioriteitssysteem met constante bedieningstijden (van meerdere slots) geanalyseerd. De lage prioriteitsbuffer wordt bestudeerd d.m.v. een wachtlijnmodel met bedieningsonderbrekingen. Wang et al. [2000] berekenen de gemiddelde vertragingstijden in een NP prioriteitswachtlijn en deze worden gebruikt voor de prestatie-analyse van een ring netwerk met meerdere kanalen. Lee [2001] en Lee et al. [2003] analyseren respectievelijk een PR en een NP prioriteitswachtlijn gebruik makend van pgf's. Tenslotte bestuderen Fiems et al. [2004] en Fiems [2004] de prestatiematen van de lage prioriteitsklasse in PR, PRD en PRI prioriteitswachtlijnen. Dit wordt gedaan a.d.h.v. een wachtlijnmodel met bedieningsonderbrekingen.

Verder zijn er een aantal discrete-tijd wachtlijnmodellen geanalyzeerd met correlatie in de aankomstprocessen van de verschillende prioriteitsklassen, nl. o.a. door Takahashi and Hashida [1991] en Walraevens et al. [2000b,c,a,d, 2001, 2002a,b, 2003a,b, 2004b]. In al deze artikels wordt gebruik gemaakt van pgf's. Takahashi and Hashida [1991] berekenen de gemiddelde vertragingstijden van de verschillende prioriteitsklassen in een NP en PR prioriteitswachtlijn, gebruik makend van vertragingstijdcycli. De analyses in deze sectie zijn gebaseerd op [Walraevens et al. 2000b,c,d, 2002a, 2003b] - waarin NP prioriteitswachtlijnen geanalyseerd zijn -, op [Walraevens et al. 2000a, 2001, 2002b, 2004b] - waarin PR prioriteitswachtlijnen bestudeerd zijn - en op [Walraevens et al. 2003b] - waarin een PRD prioriteitswachtlijn aan bod komt. De behaalde resultaten voor de PRI prioriteitswachtlijn zijn nog niet gepubliceerd.

In deze sectie bespreken we hoe we de vier prioriteitswachtlijnen met algemene bedieningstijden geanalyseerd hebben. Deze zijn in het Engelstalig gedeel-

te van het proefschrift één voor één bestudeerd. In dit Nederlandstalig gedeelte richten we ons op de verschillen tussen de analyses van deze 4 wachtlijnen alsmede op de verschillen in hun prestatiematen.

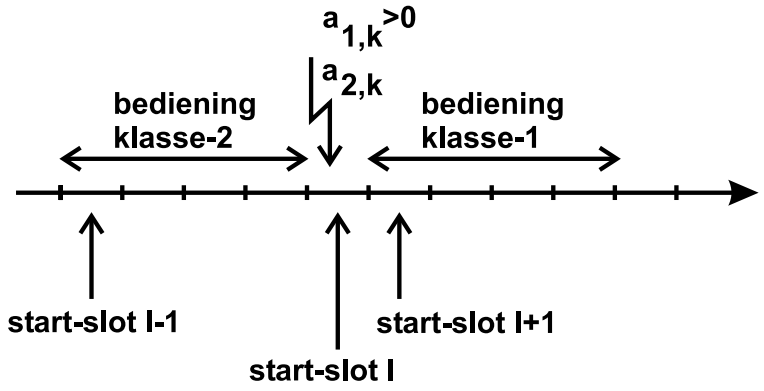
S.3.1 Bespreking van de gebruikte methodes

Aangezien de bedieningstijden van de pakketten willekeurig gedistribueerd zijn, vormen de systeembezettingen van beide klassen bij het begin van opeenvolgende slots niet langer een Markov keten. Of m.a.w., de kennis van de systeembezettingen bij het begin van een willekeurig slot (samen met het aantal aankomsten van beide klassen in dat slot) is niet genoeg informatie om de distributie van de systeembezettingen te kennen bij het begin van het volgend slot. De oplossing bestaat erin nieuwe toevalsveranderlijken te definiëren zodat een Markov keten kan opgesteld worden. Er zijn twee mogelijke manieren om een Markov keten op te bouwen en deze worden beide in de volgende paragrafen besproken.

De eerste mogelijkheid is de systeembezettingen van beide klassen te beschouwen bij het begin van specifiek gedefinieerde slots i.p.v. bij het begin van willekeurige slots. De op deze manier geconstrueerde Markov keten wordt een *ingebouwde* Markov keten (*Engels: embedded Markov chain*) genoemd, aangezien de Markov keten ingebouwd is in de keuze van de speciefieke tijdstippen. Het gebruik van deze methode beperkt zich niet tot prioriteitswachtlijnen en zelfs niet tot discrete-tijd wachtlijnen, maar wordt voor allerhande wachtlijnanalyses toegepast. Een veel gebruikte methode is bv. de systeembezettingen eerst te analyzeren op opeenvolgende *vertrektijdstippen* (als deze een Markov keten vormen) en de systeembezetting op willekeurig bepaalde tijdstippen te berekenen uit deze op vertrektijdstippen.

De tweede mogelijkheid om een Markov keten op te bouwen is het definiëren van bijkomende toevalsveranderlijken bij het begin van alle slots (naast de bufferbezettingen van beide klassen), zodanig dat de systeembezettingen van beide klassen en deze bijkomende toevalsveranderlijken bij het begin van opeenvolgende slots een Markov keten vormen. Deze techniek wordt de *bijkomende veranderlijke techniek* genoemd (*Engels: supplementary variable technique*) en wordt - volgens o.a. Chaudhry and Templeton [1983] - toegeschreven aan Kosten. Deze techniek wordt veelvuldig in discrete tijd gebruikt aangezien het grote voordeel van discrete-tijd analyses net het beschouwen van opeenvolgende slots is. In [Bruneel 1993] b.v., wordt een FIFO buffer met één klasse en algemene bedieningstijden op deze manier geanalyseerd.

In dit proefschrift hebben we beide methoden gebruikt. De eerste methode hebben we toegepast op wachtlijnen met een NP prioriteitsdiscipline, terwijl we de bijkomende veranderlijke techniek gebruikt hebben om de PR, PRD en PRI prioriteitswachtlijnen te bestuderen. Deze keuzes zijn natuurlijk niet toevallig. Het definiëren van specifieke slots zodat de systeembezettingen van



Figuur S.8: Monster van de tijdsas om de locatie van de startslots te tonen

beide klassen bij het begin van opeenvolgende van deze slots een Markov keten vormen, is een stuk eenvoudiger/natuurlijker voor het NP geval dan voor de andere gevallen. Dit komt doordat bij de NP prioriteitsdiscipline geen bedieningstijden onderbroken worden. Daardoor is het aantal slots tussen twee opeenvolgende slots waar een bediening start gelijk aan de lengte van een bedieningstijd en bevat de buffer bij het begin van deze specifiek gedefinieerde slots enkel "complete" pakketten (voor meer details zie verder). Aangezien bij de drie andere prioriteitsdisciplines de bedieningen van klasse-2 pakketten onderbroken kunnen worden, is het definiëren van zulke specifieke slots niet meer zo eenvoudig als bij het NP geval. Het definiëren van bijkomende toevalsveranderlijken bij het begin van alle slots daarentegen zodanig dat deze samen met de systeembezettingen van beide klassen bij het begin van opeenvolgende slots een Markov keten vormen, is een vrij eenvoudige zaak (zoals we verder zullen aantonen). Merk wel op dat het daarom niet noodzakelijk onmogelijk is om de verschillende prioriteitswachtlijnen ook op de andere manier te analyseren.

S.3.2 Berekeningen van de pgf's

NP prioriteitsdiscipline

We introduceren de notie van *startslots* in dit model als volgt: startslots zijn gedefinieerd als die slots bij het begin van welke de bediening van een nieuw pakket kan starten. Merk op dat uit deze definitie direct volgt dat slots bij het begin van welke het systeem geen pakketten bevat startslots zijn. In Figuur S.8 wordt een voorbeeld van een tijdsas getoond met aanduiding van de startslots.

De systeembezettingen van beide klassen bij het begin van opeenvolgende

startslots vormen een Markov keten. Inderdaad, de systeembezettingen van klasse-1 en klasse-2 bij het begin van het $l+1$-ste startslot kunnen geschreven worden als functies van de systeembezettingen van klasse-1 en klasse-2 bij het begin van het l-de startslot en de ingangsveranderlijken (specifiek: de bedieningstijd tussen beide startslots en het aantal klasse-1 en klasse-2 aankomsten gedurende deze bedieningstijd). Uitgaande van deze systeemvergelijkingen en gebruik makend van enkele wiskundige technieken (waaronder het theorema van Rouché) vinden we de gezamenlijke pgf $N(z_1, z_2)$ van de systeembezettingen van klasse-1 en klasse-2 bij het begin van een willekeurig startslot in regime. Deze gezamenlijke pgf $N(z_1, z_2)$ is vervolgens het uitgangspunt van alle verdere berekeningen.

De gezamenlijke pgf $U(z_1, z_2)$ van de klasse-1 en klasse-2 systeembezettingen bij het begin van een willekeurig slot in regime wordt hier vooreerst uit berekend en is gegeven door

$$\begin{aligned} U(z_1, z_2) =& (1-\rho_T)\frac{E_1(z_1, z_2)(z_1-1)}{z_1 - E_1(z_1, z_2)} + (1-\rho_T) \\ & \times \frac{(A(Y_1(z_2), z_2) - 1)\left\{\begin{array}{l} z_1 E_2(z_1, z_2)(E_1(z_1, z_2) - 1) \\ +z_2 E_1(z_1, z_2)(1 - E_2(z_1, z_2)) \\ +z_1 z_2 (E_2(z_1, z_2) - E_1(z_1, z_2)) \end{array}\right\}}{(A(z_1, z_2) - 1)(z_1 - E_1(z_1, z_2))(z_2 - Y_2(z_2))}, \end{aligned} \tag{S.27}$$

met

$$E_j(z_1, z_2) \triangleq S_j(A(z_1, z_2)) \tag{S.28}$$

$$Y_j(z) \triangleq E_j(Y_1(z), z). \tag{S.29}$$

Net als in de analyse in de vorige sectie is deze pgf $U(z_1, z_2)$ niet volledig expliciet gekend, aangezien $Y_1(z)$ impliciet gedefinieerd is. Hieruit kunnen dan de marginale pgf's van de totale, klasse-1 en klasse-2 systeembezettingen berekend worden.

Vervolgens worden de vertragingstijden van klasse-1 en klasse-2 pakketten (afzonderlijk) bestudeerd. Dit gebeurt door de vertragingstijden van een willekeurig klasse-1 of willekeurig klasse-2 pakket uit te drukken i.f.v. de systeembezetting bij het begin van het laatste startslot voor zijn aankomstslot en deze te z-transformeren. Bij de klasse-2 vertragingstijd gebruiken we opnieuw de notie van de fundamentele periode. Bij algemene bedieningstijden maken we echter onderscheid tussen twee soorten fundamentele periodes, nl. *fundamentele periodes geïnitieerd door een klasse-1 pakket* en *fundamentele periodes geïnitieerd door een klasse-2 pakket*. Een fundamentele periode geïnitieerd door een klasse-1 pakket wordt gedefinieerd als de periode (uitgedrukt in een aantal slots) die nodig is om het aantal klasse-1 pakketten - vanaf het begin van de bedieningstijd van het klasse-1 pakket - te verminderen met 1. Een fundamentele periode geïnitieerd door een klasse-2 pakket start bij het begin van de

bedieningstijd van dit klasse-2 pakket en eindigt wanneer de bediening van een volgend klasse-2 pakket kan starten. Deze twee soorten fundamentele periodes zijn verschillend gedistribueerd als de bedieningstijden van klasse-1 en klasse-2 pakketten een verschillende distributie hebben. Daarom was het in het vorige model - constante bedieningstijden van één slot - niet nodig om het onderscheid te maken tussen deze twee types fundamentele periodes. Uiteindelijk verkrijgen we de volgende pgf's voor de klasse-1 en klasse-2 vertragingstijden in de NP prioriteitswachtlijn:

$$D_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1} \left(\frac{1-\rho_T}{1-\rho_1} + \frac{\rho_2}{1-\rho_1} \frac{S_2(z)-1}{\mu_2(z-1)} \right) \tag{S.30}$$

$$D_2(z) = \frac{1-\rho_T}{\lambda_2} \frac{S_2(z)(A(V_1(z), V_2(z)) - A_1(V_1(z)))}{zA_1(V_1(z)) - A(V_1(z), V_2(z))} \frac{zA_1(V_1(z)) - 1}{V_2(z) - 1}, \tag{S.31}$$

met

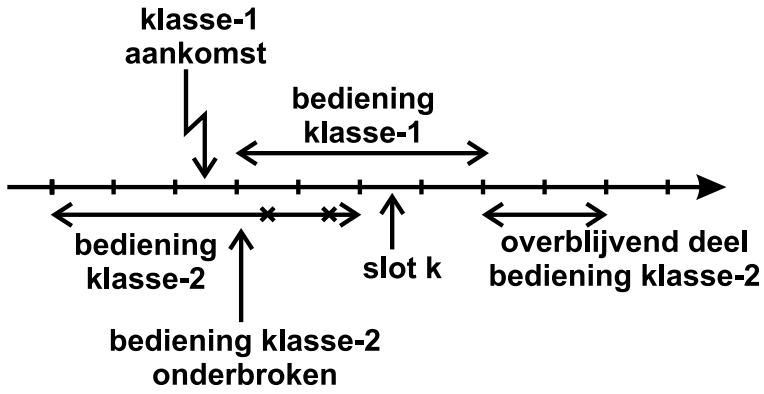
$$V_j(z) \triangleq S_j(zA_1(V_1(z))), \tag{S.32}$$

$j = 1, 2$, de pgf van een fundamentele periode geïnitieerd door een klasse-j pakket. Merk op dat $V_1(z)$ - net als $Y_1(z)$ - enkel impliciet gedefinieerd is. De pgf van de vertragingstijd van een willekeurig pakket is dan een gewogen som van $D_1(z)$ en $D_2(z)$ (met respectievelijke gewichten λ_1/λ_T en λ_2/λ_T).

PR prioriteitsdiscipline

Zoals al vermeld, maken we voor de analyse van de PR prioriteitswachtlijn gebruik van de bijkomende veranderlijke techniek. We analyseren echter niet onmiddellijk een wachtlijn met willekeurige bedieningstijden, maar starten met een vereenvoudigd model dat we gaandeweg veralgemenen.

Eerst veronderstellen we de bedieningstijden geometrisch verdeeld. Het specifieke aan de geometrische distributie is het feit dat ze de *geheugenloze eigenschap* bezit. Toegepast op de bedieningstijden wil dit specifiek zeggen dat als we een slot uitkiezen waarin een bedieningstijd aan de gang is, dat het residuele deel van deze bedieningstijd niet afhangt van het aantal slots dat het pakket al bediend is. M.a.w. de kans dat een pakket in bediening nog minstens een slot bediening nodig heeft, kan beschreven worden door één parameter, onafhankelijk van hoelang dat pakket al in bediening is. Daardoor vormen de systeembezettingen van beide klassen bij het begin van opeenvolgende slots een Markov keten en moeten er voor dit vereenvoudigd model geen bijkomende toevalsveranderlijken gedefinieerd worden. Merk op dat de parameters van de geometrische distributies van de klasse-1 en klasse-2 bedieningstijden verschillend kunnen zijn. De gezamenlijke pgf $U(z_1, z_2)$ van de systeembezettingen van beide klassen kan dan berekend worden.



Figuur S.9: Monster van de tijdsas om de prioriteitsdiscipline en de bijkomende veranderlijken te illustreren

Dit eerste model breiden we dan vooreerst uit naar algemene bedieningstijden voor de klasse-1 pakketten. In dit geval vormen de systeembezettingen van beide klassen bij het begin van opeenvolgende slots niet langer een Markov keten. Daarom definiëren we een bijkomende veranderlijke, namelijk, het overblijvend aantal slots dat het klasse-1 pakket nog in bediening is alvorens het systeem te verlaten. Deze toevalsveranderlijke wordt de *residuele bedieningstijd van klasse-1* genoemd. Aangezien de klasse-2 bedieningstijden in dit tweede model nog steeds geometrisch verdeeld zijn, is deze nieuwe bijkomende stochastische veranderlijke - samen met de systeembezettingen van beide klassen - voldoende voor een Markoviaanse beschrijving van het systeem. In Figuur S.9 hebben we een voorbeeld van de tijdsas gegeven om de concepten te illustreren: een klasse-1 pakket komt aan tijdens een klasse-2 bediening. Deze laatste wordt onderbroken en later voortgezet. De residuele bedieningstijd van klasse-1 bij het begin van slot k is gelijk aan twee slots. De gezamenlijke pgf van de residuele bedieningstijd van klasse-1, de systeembezetting van klasse-1 en de systeembezetting van klasse-2 bij het begin van een willekeurig slot in regime is dan berekend. Uit deze driedimensionale pgf kunnen alle andere pgf's van belang afgeleid worden.

De laatste uitbreiding is het veralgemenen naar algemene bedieningstijden voor beide klassen. In dit geval wordt - naast de residuele bedieningstijd van klasse-1 - nog een veranderlijke toegevoegd om een Markoviaanse beschrijving van het systeem te bekomen, nl., het overblijvende aantal slots van de bedieningstijd van het oudste klasse-2 pakket. (Het oudste klasse-j pakket in het systeem bij het begin van een zeker slot is gedefinieerd als dit klasse-j pakket dat - van alle klasse-j pakketten op dat tijdstip aanwezig - het eerst aankwam.) Deze nieuw gedefinieerde toevalsveranderlijke wordt de *residuele bedieningstijd van klasse-2* genoemd. De systeembezettingen en de residuele bedieningstijden van beide klassen bij het begin van opeenvolgende slots vor-

men dan een Markov keten. Bekijken we opnieuw als voorbeeld de tijdsas van Figuur S.9: de residuele bedieningstijd van klasse-2 bij het begin van slot k is gelijk aan twee slots. De vierdimensionale pgf $P(x_1, z_1, x_2, z_2)$ van de residuele bedieningstijd van klasse-1, de systeembezetting van klasse-1, de residuele bedieningstijd van klasse-2 en de systeembezetting van klasse-2 wordt berekend.

Voor de drie besproken modellen kunnen dan alle verdere pgf's van belang afgeleid worden uit de berekende gezamenlijke pgf's. In wat volgt, bespreken we enkel de verdere analyse van het meest uitgebreide model, nl., algemene bedieningstijden voor beide klassen.

Uitgaande van de uitdrukking voor de vierdimensionale pgf $P(x_1, z_1, x_2, z_2)$ wordt $U(z_1, z_2)$ verkregen als

$$\begin{aligned} U(z_1, z_2) =& P(1, z_1, 1, z_2) && \text{(S.33)} \\ =& (1-\rho_T)\frac{Y_2(z_2)(z_2-1)}{z_2 - Y_2(z_2)} && \text{(S.34)} \\ & \times \left[1 + z_1 \frac{(A(z_1,z_2) - A(Y_1(z_2),z_2))(E_1(z_1,z_2)-1)}{A(Y_1(z_2),z_2)(A(z_1,z_2)-1)(z_1 - E_1(z_1,z_2))}\right], \end{aligned}$$

met

$$Y_j(z) \triangleq E_j(Y_1(z), z). \tag{S.35}$$

Merk op dat deze $Y_j(z)$ identiek gedefinieerd is als in de NP prioriteitswachtlijn. Uit $U(z_1, z_2)$ kunnen dan de marginale pgf's van de totale, klasse-1 en klasse-2 systeembezettingen berekend worden.

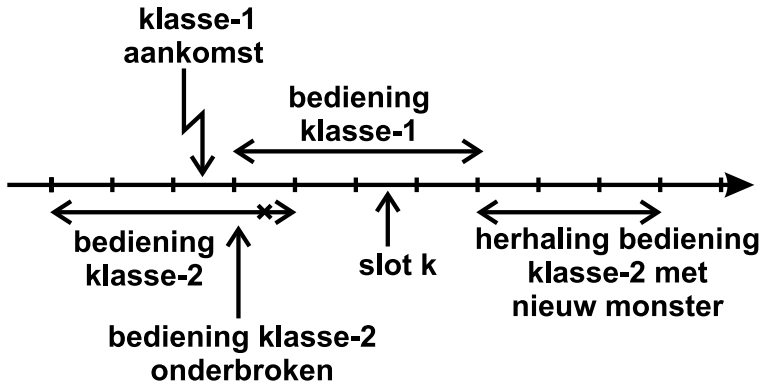
De vertragingstijden van klasse-1 en klasse-2 pakketten zijn vervolgens geanalyseerd. Dit gebeurt door de vertragingstijden van een willekeurig klasse-1 of willekeurig klasse-2 pakket uit te drukken i.f.v. de veranderlijken in de Markoviaanse beschrijving bij het begin van het aankomstslot van het pakket en deze te z-transformeren. Daardoor verkrijgen we $D_1(z)$ en $D_2(z)$ als functies van $P(.,.,.,.)$. Bij de klasse-2 vertragingstijd gebruiken we dan opnieuw de notie van de fundamentele periodes. Deze zijn identiek gedistibueerd als in de NP prioriteitswachtlijn. We verkrijgen

$$D_1(z) = \frac{1-\rho_1}{\lambda_1}\frac{S_1(z)(z-1)}{z - A_1(S_1(z))}\frac{A_1(S_1(z))-1}{S_1(z)-1} \tag{S.36}$$

$$D_2(z) = \frac{1-\rho_T}{\lambda_2}\frac{V_2(z)(zA_1(V_1(z))-1)}{A_1(V_1(z))(V_2(z)-1)}\frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))}, \tag{S.37}$$

met de pgf van een fundamentele periode geïnitieerd door een klasse-j pakket nog steeds gegeven door

$$V_j(z) \triangleq S_j(zA_1(V_1(z))), \tag{S.38}$$



Figuur S.10: Monster van de tijdsas in het PRD geval om de prioriteitsdiscipline en de bijkomende veranderlijken te illustreren

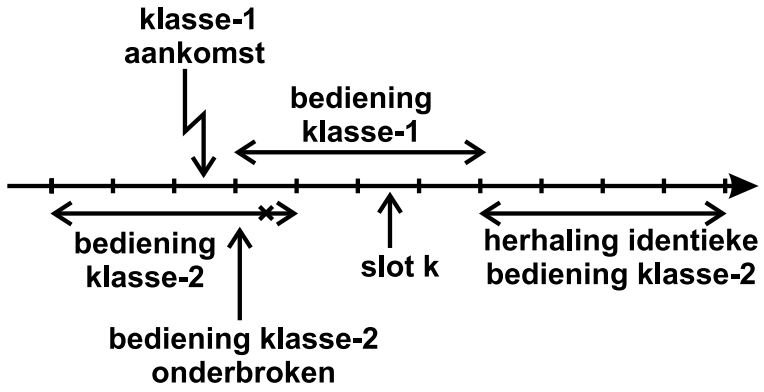
$j = 1, 2$. De pgf van de vertragingstijd van een willekeurig pakket is dan een gewogen som van de $D_j(z)$ (met gewichten λ_j/λ_T respectievelijk).

PRD en PRI prioriteitsdisciplines

In de analyses van de PRD en PRI prioriteitswachtlijnen - i.e. de preëmptieve prioriteitswachtlijnen met herhaling - maken we opnieuw gebruik van de bijkomende veranderlijke techniek.

In het geval dat een nieuw monster van de klasse-2 bedieningstijden gekozen wordt bij herhaling, i.e. in het PRD-geval, definiëren we de *residuele bedieningstijd van het pakket in bediening bij het begin van een slot* als de bijkomende veranderlijke om een Markoviaanse beschrijving van het systeem te bekomen. Deze toevalsveranderlijke is gedefinieerd als het overblijvende deel van het pakket (dat zowel van klasse-1 als klasse-2 kan zijn) in bediening bij het begin van een willekeurig slot. Figuur S.10 geeft een voorbeeld van een tijdsas weer. De onderbroken bediening van het klasse-2 pakket wordt later herhaald, maar de lengte kan veranderen (in het voorbeeldje is deze lengte vier slots bij de eerste poging en drie slots bij de tweede). Bij het begin van slot k is de residuele bediening van het pakket in bediening 2 slots. De gezamenlijke pgf $P(x, z_1, z_2)$ van de residuele bedieningstijd van het pakket in bediening en de systeembezettingen van klasse-1 en klasse-2 bij het begin van een willekeurig slot in regime is dan berekend.

In het geval van een PRI prioriteitsdiscipline, i.e., het geval waarbij een onderbroken klasse-2 bedieningstijd van dezelfde lengte blijft bij herhalingen, is de residuele bedieningstijd van het pakket in bediening niet meer voldoende als bijkomende veranderlijke om een Markov keten te vormen. Daarom definiëren we naast deze eerste bijkomende veranderlijke een tweede, nl. de *(vol-*



Figuur S.11: Monster van de tijdsas in het PRI geval om de prioriteitsdiscipline en de bijkomende veranderlijken te illustreren

ledige) bedieningstijd van het oudste klasse-2 pakket. Figuur S.11 geeft opnieuw een voorbeeld van een tijsdas. De onderbroken bediening van het klasse-2 pakket wordt later herhaald en de lengte blijft identiek. De nieuw gedefinieerde bijkomende veranderlijke is bij het begin van alle slots in de figuur gelijk aan vier slots. We verkrijgen dan een uitdrukking voor de gezamenlijke pgf $P(x, z_1, y_2, z_2)$ van de residuele bedieningstijd van het pakket in bediening, de systeembezetting van klasse-1, de bedieningstijd van het klasse-2 pakket langst in het systeem en de systeembezetting van klasse-2 bij het begin van een willekeurig slot in regime na vrij extensieve berekeningen.

Uitgaande van de uitdrukkingen voor $P(x, z_1, z_2)$ en $P(x, z_1, y_2, z_2)$ in respectievelijk de PRD en PRI prioriteitswachtlijn berekenen we $U(z_1, z_2)$ - de gezamenlijke pgf van de systeembezettingen van klasse-1 en klasse-2 bij het begin van een willekeurig slot in regime - voor beide prioriteitsdisciplines. We verkrijgen

$$\begin{aligned} U(z_1, z_2) =& (1 - \rho_{T,eff}) \frac{Y_2(z_2)(z_2 - 1)}{z_2 - Y_2(z_2)} \\ & \times \left\{ 1 + z_1 \frac{(A(z_1, z_2) - A(Y_1(z_2), z_2))(E_1(z_1, z_2) - 1)}{A(Y_1(z_2), z_2)(A(z_1, z_2) - 1)(z_1 - E_1(z_1, z_2))} \right\}, \end{aligned} \tag{S.39}$$

in beide gevallen met

$$\rho_{T,eff} \triangleq \rho_1 + \lambda_2 \mu_{2,eff} \tag{S.40}$$

$$\mu_{2,eff} \triangleq \frac{A_1(0)(1 - S_2(A_1(0)))}{S_2(A_1(0))(1 - A_1(0))} \tag{S.41}$$

$$Y_1(z) \triangleq E_1(Y_1(z), z) \tag{S.42}$$

$$Y_2(z) \triangleq \frac{(1-A(0,z))A(Y_1(z),z)E_2(0,z)}{(A(Y_1(z),z)-A(0,z))E_2(0,z)-A(0,z)(A(Y_1(z),z)-1)}, \quad \text{(S.43)}$$

in het geval van PRD en

$$\rho_{T,eff} \triangleq \rho_1 + \lambda_2 \mu_{2,eff} \quad \text{(S.44)}$$

$$\mu_{2,eff} \triangleq \frac{S_2(1/A_1(0))-1}{1/A_1(0)-1} \quad \text{(S.45)}$$

$$Y_1(z) \triangleq E_1(Y_1(z),z) \quad \text{(S.46)}$$

$$Y_2(z) \triangleq \sum_{i=1}^{\infty} \frac{s_2(i)(1-A(0,z))A(Y_1(z),z)A(0,z)^i}{(A(Y_1(z),z)-A(0,z))A(0,z)^i - A(0,z)(A(Y_1(z),z)-1)}, \quad \text{(S.47)}$$

in het geval van PRI. Merk op dat alhoewel $Y_1(z)$ en $Y_2(z)$ nog dezelfde 'rol' spelen als in de NP en PR wachtlijnen, de uitdrukkingen voor $Y_2(z)$ in deze gevallen een stuk ingewikkelder zijn. Dit komt doordat de bedieningstijden herhaald moeten worden. Zo is ook de *effectieve belasting* $\rho_{T,eff}$ niet langer gelijk aan de aankomstbelasting ρ_T - zoals het geval was in de NP en PR prioriteitswachtlijnen - aangezien de laatste geen rekening houdt met de mogelijke herhalingen van de klasse-2 bedieningen. Uit de uitdrukkingen voor $U(z_1,z_2)$ kunnen dan opnieuw de marginale pgf's van de totale, klasse-1 en klasse-2 systeembezettingen berekend worden.

Tenslotte berekenen we de vertragingstijden van klasse-1 en klasse-2 pakketten voor de PRD en PRI prioriteitswachtlijnen. Dit gebeurt opnieuw door de vertragingstijden van een willekeurig klasse-1 en willekeurig klasse-2 pakket uit te drukken i.f.v. de veranderlijken in de Markoviaanse beschrijving bij het begin van het aankomstslot van het pakket en deze te z-transformeren. Bij de klasse-2 vertragingstijd gebruiken we opnieuw de notie van fundamentele periodes. Deze geïnitieerd door een klasse-2 pakket zijn echter in dit geval een stuk ingewikkelder om te analyseren (opnieuw wegens de mogelijke herhalingen van klasse-2 bedieningstijden). Uiteindelijk verkrijgen we

$$D_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{S_1(z)(z-1)}{z-A_1(S_1(z))} \frac{A_1(S_1(z))-1}{S_1(z)-1}, \quad \text{(S.48)}$$

voor de pgf van de klasse-1 vertragingstijd in zowel de PRD als de PRI prioriteitswachtlijn. Merk op dat dit dezelfde uitdrukking is als voor $D_1(z)$ in de PR prioriteitswachtlijn, aangezien klasse-1 pakketten geen hinder ondervinden van klasse-2 pakketten bij de drie preëmptieve prioriteitsdisciplines en het dus wat betreft de klasse-1 systeemkarakteristieken niet uitmaakt of een klasse-2 bediening voortgezet of herhaald wordt na een onderbreking. De pgf van de klasse-2 vertragingstijd wordt gegeven door

$$D_2(z) = \frac{1-\rho_{T,eff}}{\lambda_T} \frac{V_2(z)}{A_1(V_1(z))} \left\{ \frac{(zA_1(V_1(z))-1)(A(0,V_2(z))-A_1(0))}{(V_2(z)-1)(A_1(0)z-A(0,V_2(z)))} \right.$$

$$+\frac{(zA_1(V_1(z))-A(Y_1(V_2(z)),V_2(z)))C(z)}{A(Y_1(V_2(z)),V_2(z))} \times\frac{(A_1(0)A(V_1(z),V_2(z))-A_1(V_1(z))A(0,V_2(z)))(z-1)}{(A_1(0)z-A(0,V_2(z)))(zA_1(V_1(z))-A(V_1(z),V_2(z)))}\Bigg\}, \quad \text{(S.49)}$$

in beide prioriteitswachtlijnen met

$$C(z) \triangleq \frac{Y_2(V_2(z))}{V_2(z)-Y_2(V_2(z))} \quad \text{(S.50)}$$

$$V_1(z) \triangleq S_1(zA_1(V_1(z))) \quad \text{(S.51)}$$

$$V_2(z) \triangleq \frac{(1-A_1(0)z)A_1(V_1(z))S_2(A_1(0)z)}{(A_1(V_1(z))-A_1(0))S_2(A_1(0)z)-A_1(0)(zA_1(V_1(z))-1)}, \quad \text{(S.52)}$$

in het PRD geval en met

$$C(z) \triangleq \frac{\sum_{i=1}^{\infty} s_2(i)(V_{2,i}(z)-1)Y_{2,i}(V_2(z))}{(V_2(z)-Y_2(V_2(z)))(V_2(z)-1)} \quad \text{(S.53)}$$

$$V_1(z) \triangleq S_1(zA_1(V_1(z))) \quad \text{(S.54)}$$

$$V_{2,i}(z) \triangleq \frac{(1-A_1(0)z)A_1(V_1(z))(A_1(0)z)^i}{(A_1(V_1(z))-A_1(0))(A_1(0)z)^i-A_1(0)(zA_1(V_1(z))-1)} \quad \text{(S.55)}$$

$$V_2(z) \triangleq \sum_{i=1}^{\infty} s_2(i)V_{2,i}(z) \quad \text{(S.56)}$$

$$Y_{2,i}(z) \triangleq \frac{(1-A(0,z))A(Y_1(z),z)A(0,z)^i}{(A(Y_1(z),z)-A(0,z))A(0,z)^i-A(0,z)(A(Y_1(z),z)-1)}, \quad \text{(S.57)}$$

in het PRI geval.

Het is duidelijk dat deze uitdrukkingen een stuk gecompliceerder zijn dan de (overeenkomstige) uitdrukkingen voor de NP en PR prioriteitswachtlijnen. Dit is enerzijds te wijten aan de herhalingen van de onderbroken klasse-2 bedieningstijden en anderzijds aan de mogelijke correlatie tussen klasse-1 en klasse-2 aankomsten in een slot. De pgf van de vertragingstijd van een willekeurig pakket is opnieuw een gewogen som van $D_1(z)$ en $D_2(z)$ met respectievelijke gewichten λ_1/λ_T en λ_2/λ_T.

S.3.3 Berekening prestatiematen

Uit de behaalde uitdrukkingen voor de pgf's kunnen wederom de momenten en staartprobabiliteiten berekend worden van de verschillende toevalsveranderlijken. De procedures voor de berekeningen van momenten en staartprobabiliteiten zijn heel gelijkaardig aan die uit de vorige sectie en we gaan er dan ook niet dieper op in, maar verwijzen naar subsectie S.2.2 voor meer details omtrent het berekenen van deze prestatiematen.

Bij het berekenen van staartprobabiliteiten in de PRI prioriteitswachtlijn is er echter een extra moeilijkheid, nl. het optreden van de oneindige sommen in de uitdrukkingen voor $Y_2(z)$ en $V_2(z)$ (respectievelijk uitdrukkingen (S.47) en (S.56)). Het lijkt geen triviaal probleem te zijn om de dominante singulariteit en het gedrag van de pgf's in de buurt van deze singulariteit te bepalen. Daarom hebben we in het proefschrift geen staartprobabiliteiten kunnen berekenen in het PRI-geval.

S.3.4 Numerieke voorbeelden

In deze paragraaf zullen we kort de invloed van enkele parameters op de prestatiematen tonen. We concentreren ons op het schakelelement met uitgangsbuffers zoals eerder vermeld (zie Figuur S.1). De gezamenlijke pgf van het aantal klasse-j aankomsten ($j = 1, 2$) aan een willekeurige uitgangsbuffer is gegeven door (S.6). In het vervolg definiëren we α als de fractie klasse-1 belasting van de totale belasting (dus gelijk aan ρ_1/ρ_T). We veronderstellen dat N - het aantal ingangen van het schakelelement - gelijk is aan 16.

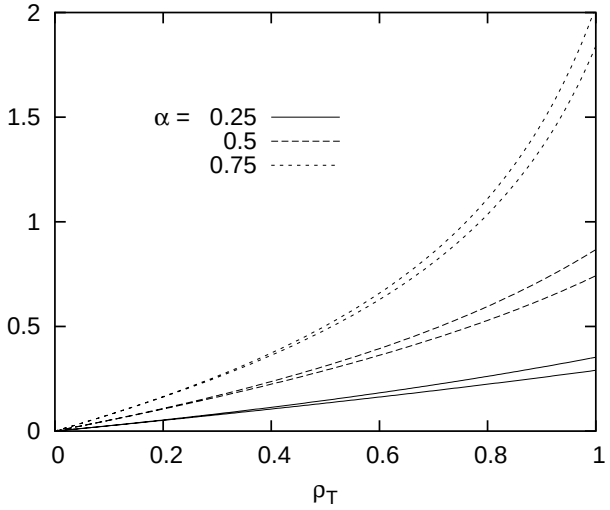
We zullen ons in deze subsectie vooral richten op de verschillen tussen de prestatiematen in de verschillende prioriteitswachtlijnen. Eerst zullen we de klasse-1 en klasse-2 prestatiematen in de NP en PR prioriteitssystemen vergelijken. Vervolgens bekijken we wat de invloed van herhalingen is - t.o.v. voortzetting - op de klasse-2 prestatiematen bij de preëmptieve prioriteitsdisciplines.

Vergelijking van de NP en PR prioriteitsdisciplines

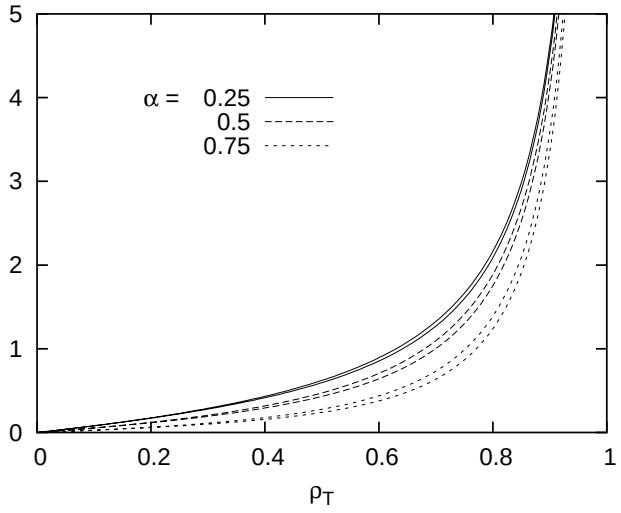
De bedieningstijden zijn constant verondersteld doorheen deze paragraaf.

In Figuren S.12 en S.13 tonen we respectievelijk de gemiddelde klasse-1 en klasse-2 systeembezettingen als functies van de totale belasting met $\mu_1 = \mu_2 = 2$ en $\alpha = 0.25, 0.5$ en 0.75. In beide figuren, zijn de curven voor de NP en PR prioriteitsdiscipline uitgezet. De gemiddelde klasse-1 systeembezetting is groter in het geval van de NP prioriteitsdiscipline. Het omgekeerde geldt voor de gemiddelde klasse-2 systeembezetting. Dit is logisch, aangezien in de NP prioriteitswachtlijn aankomende klasse-1 pakketten een klasse-2 bediening niet onderbreken en dus hinder ondervinden van het klasse-2 verkeer, terwijl dit in de PR prioriteitswachtlijn niet het geval is. In het PR geval kunnen de klasse-2 bedieningen meermaals onderbroken worden waardoor ze langer in de buffer verblijven en daardoor is de gemiddelde klasse-2 systeembezetting in dit geval groter.

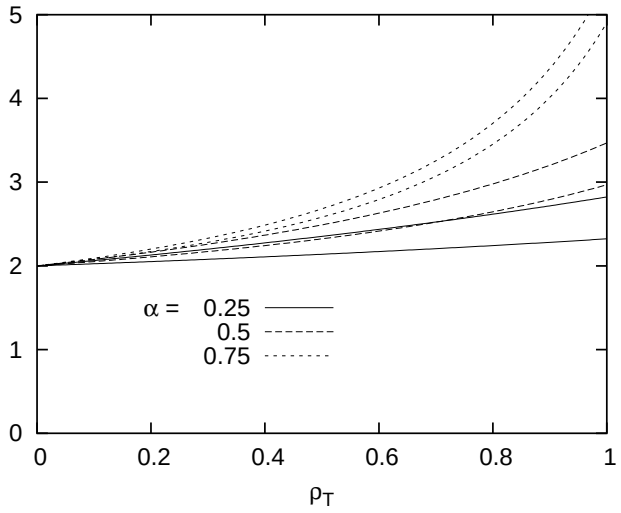
Gelijkaardige verschillen in de invloeden van de NP en PR prioriteitsdisciplines kunnen gezien worden in Figuren S.14 en S.15, waarin respectievelijk de gemiddelde klasse-1 en klasse-2 vertragingstijden als functies van de totale belasting getoond zijn, voor dezelfde parameters als in de twee vorige



Figuur S.12: Gemiddelde klasse-1 systeembezetting versus de totale belasting voor de NP (bovenste curven) en PR (onderste curven) prioriteitsdisciplines ($\mu_1 = \mu_2 = 2$)



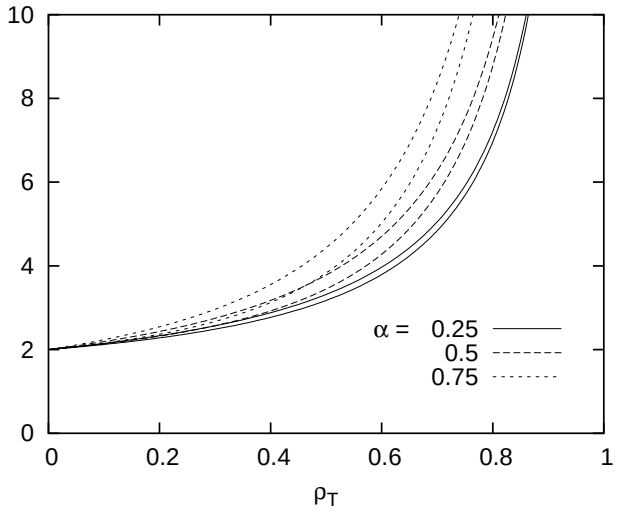
Figuur S.13: Gemiddelde klasse-2 systeembezetting versus de totale belasting voor de NP (onderste curven) en PR (bovenste curven) prioriteitsdisciplines ($\mu_1 = \mu_2 = 2$)



Figuur S.14: Gemiddelde klasse-1 vertragingstijd versus de totale belasting voor de NP (bovenste curven) en PR (onderste curven) prioriteitsdisciplines ($\mu_1 = \mu_2 = 2$)

figuren. Uit deze figuren blijkt dat de verschillen significant kunnen zijn. Zo kan b.v. voor $\alpha = 0.25$ de gemiddelde klasse-1 vertragingstijd in het NP geval ongeveer 25% groter worden dan in het PR geval.

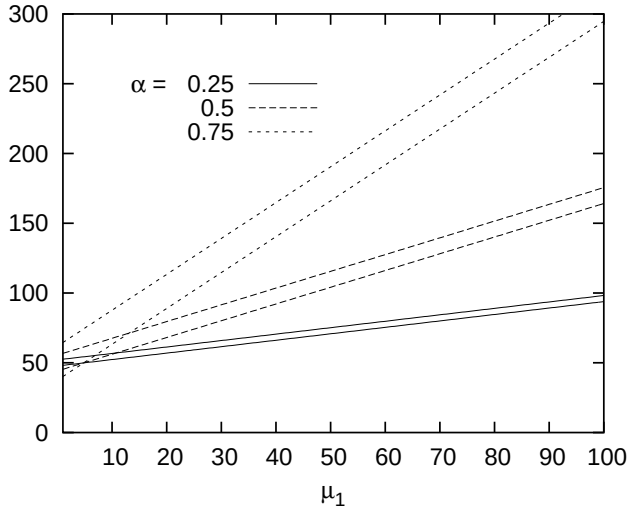
Vervolgens hebben we de invloed van de lengte van de bedieningstijden op de gemiddelde systeembezettingen en vertragingstijden bekeken. In Figuur S.16 tonen we de gemiddelde klasse-2 vertragingstijd als functie van de gemiddelde klasse-1 bedieningstijd voor zowel de NP als de PR prioriteitsdisciplines, met $\mu_2 = 20$ slots, $\rho_T = 0.75$ en $\alpha = 0.25, 0.5$ en 0.75. Het is duidelijk dat de gemiddelde klasse-2 vertragingstijden groter worden naarmate de klasse-1 pakketten langer worden. Dit is een direct gevolg van de prioriteitsdisciplines. Om de invloed van de NP discipline op de gemiddelde klasse-1 vertragingstijd aan te tonen, zetten we in Figuur S.17 de gemiddelde klasse-1 vertragingstijd uit als functie van de klasse-2 bedieningstijden voor $\mu_1 = 20$ en de andere parameters identiek aan deze uit de vorige figuur, voor zowel de PR als de NP prioriteitsdisciplines. Aangezien de klasse-1 vertragingstijd in het PR geval onafhankelijk is van het klasse-2 verkeer, is in dit geval de gemiddelde klasse-1 vertragingstijd onafhankelijk van μ_2. In het NP geval daarentegen, zien we dat de gemiddelde klasse-1 vertragingstijd stijgt met μ_2. Dit is omdat aankomende klasse-1 pakketten (gemiddeld gezien) langer moeten wachten wanneer een klasse-2 pakket in bediening is (de residuele bedieningstijd van dit klasse-2 pakket bij aankomst van klasse-1 pakketten zal gemiddeld gezien groter zijn naarmate de klasse-2 bedieningstijden zelf groter zijn). Het is duidelijk dat voor lange klasse-2 pakketten de klasse-1 vertragingstijd (te) hoog



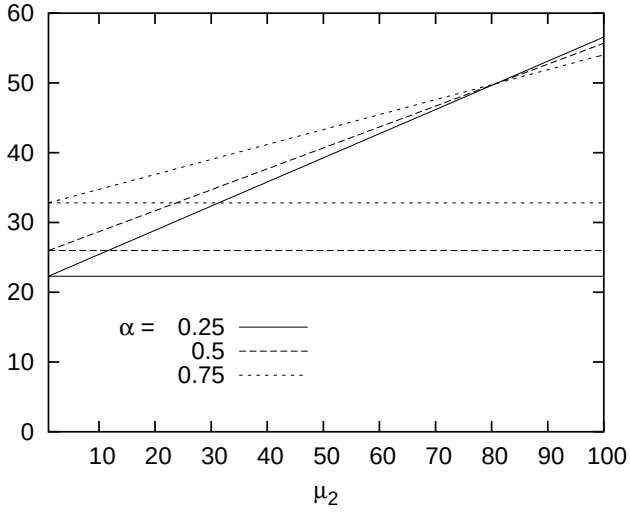
Figuur S.15: Gemiddelde klasse-2 vertragingstijd versus de totale belasting voor de NP (onderste curven) en PR (bovenste curven) prioriteitsdisciplines ($\mu_1 = \mu_2 = 2$)

kan oplopen.

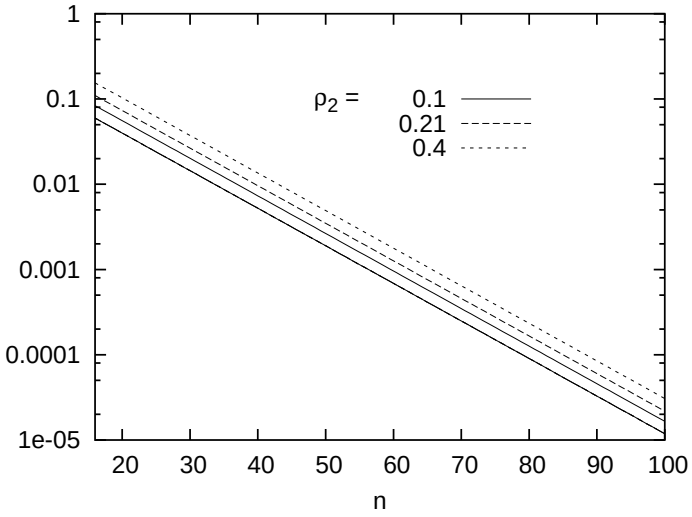
Vervolgens vergelijken we de staartprobabiliteiten van de klasse-1 en klasse-2 vertragingstijden in het NP en PR geval. Het is zo dat de dominante singulariteiten van $D_1(z)$ en $D_2(z)$ - de pgf's van de klasse-1 en klasse-2 vertragingstijden respectievelijk - die een rol spelen in het overeenkomstig staartgedrag gelijk zijn voor de beide prioriteitsdisciplines. De hellingen van de curves worden nu net uitsluitend bepaald door deze singulariteiten en dus zullen de hellingen voor de beide disciplines gelijk zijn. Figuren S.18 en S.19 tonen de staartprobabiliteiten van respectievelijk de klasse-1 en klasse-2 vertragingstijden voor beide types prioriteitsdisciplines. De klasse-1 belasting is voor alle curven gelijk aan 0.4 terwijl de klasse-2 belasting de waarden 0.1, (ongeveer) 0.21 en 0.4 aanneemt. De bedieningstijden van alle pakketten zijn gelijk aan 16. In Figuur S.18 is de onderste curve die voor de klasse-1 staartprobabiliteiten in het PR geval. Aangezien voor deze discipline de klasse-1 vertragingstijden onafhankelijk zijn van de klasse-2 karakteristieken, krijgen we inderdaad dezelfde curve voor de verschillende waarden van de klasse-2 belasting. In Figuur S.19 zijn de waarden van ρ_2 zodanig gekozen dat opnieuw de 3 types staartprobabiliteiten getoond zijn: nl. niet-geometrische gedrag treedt op voor $\rho_2 = 0.1$, transitiegedrag voor $\rho_2 = 0.21$ en geometrisch gedrag voor $\rho_2 = 0.4$. Verder kan uit deze twee figuren geconcludeerd worden dat het type van prioriteitsdiscipline een (niet te verwaarlozen) rol speelt in de staartprobabiliteiten van de vertragingstijden van beide klassen.



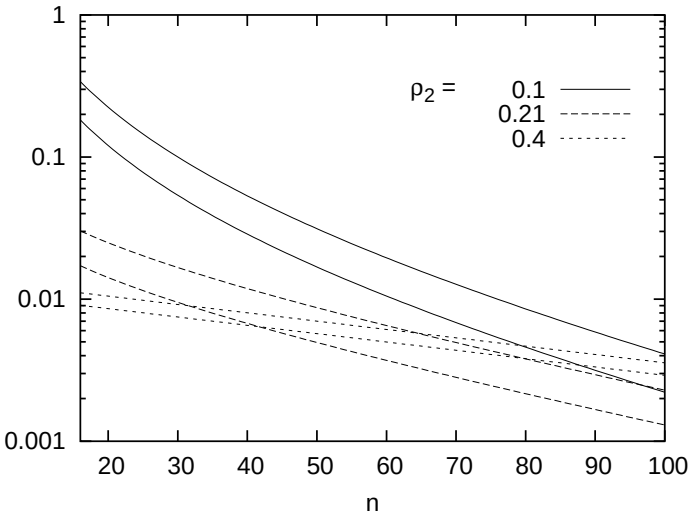
Figuur S.16: Gemiddelde klasse-2 vertragingstijd versus de gemiddelde klasse-1 bedieningstijden voor de NP (onderste curven) en PR (bovenste curven) prioriteitsdisciplines ($\rho_T = 0.75, \mu_2 = 20$)



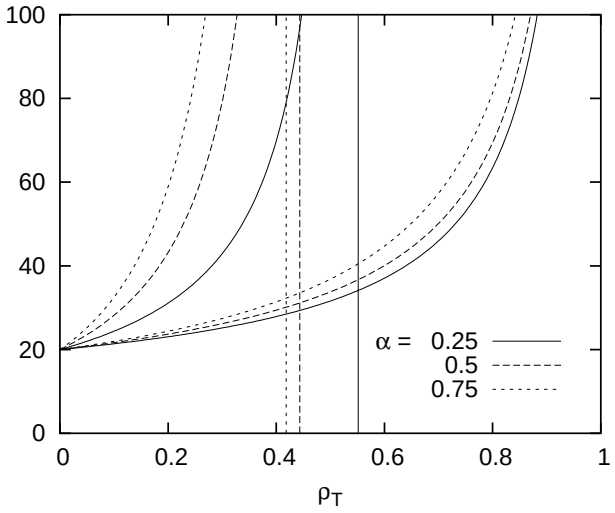
Figuur S.17: Gemiddelde klasse-1 vertragingstijd versus de gemiddelde klasse-2 bedieningstijden voor de NP (bovenste curven) en PR (onderste curven) prioriteitsdisciplines ($\rho_T = 0.75, \mu_1 = 20$)



Figuur S.18: Staartgedrag van de klasse-1 vertragingstijden voor verschillende klasse-2 belastingen voor zowel de NP (bovenste curven) als de PR (onderste curve) prioriteitsdisciplines ($\rho_1 = 0.4, \mu_1 = \mu_2 = 16$)



Figuur S.19: Staartgedrag van de klasse-2 vertragingstijden voor verschillende klasse-2 belastingen voor zowel de NP (onderste curven) als de PR (bovenste curven) prioriteitsdisciplines ($\rho_1 = 0.4, \mu_1 = \mu_2 = 16$)

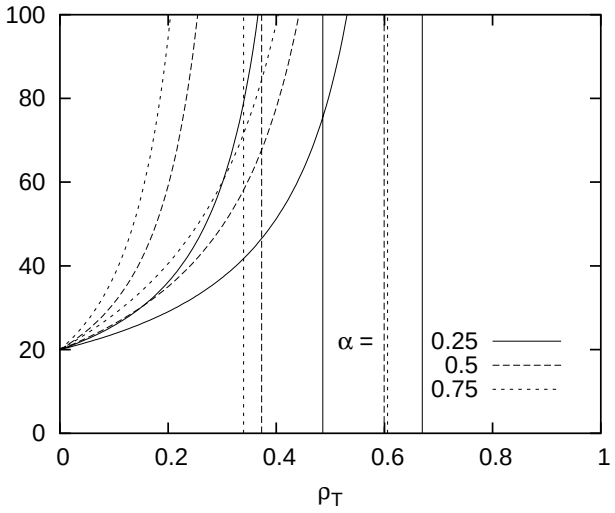


Figuur S.20: Gemiddelde klasse-2 vertragingstijd versus de totale belasting voor de PR (onderste curven) en PRD en PRI (bovenste curven) prioriteitsdisciplines voor constante bedieningstijden ($\mu_1 = 2, \mu_2 = 20$)

Vergelijking van de PR, PRD en PRI prioriteitsdisciplines

Aangezien in al deze preëmptieve disciplines de klasse-1 veranderlijken identiek zijn, focussen we enkel op de klasse-2 prestatie (en meer specifiek op de gemiddelde klasse-2 vertragingstijden).

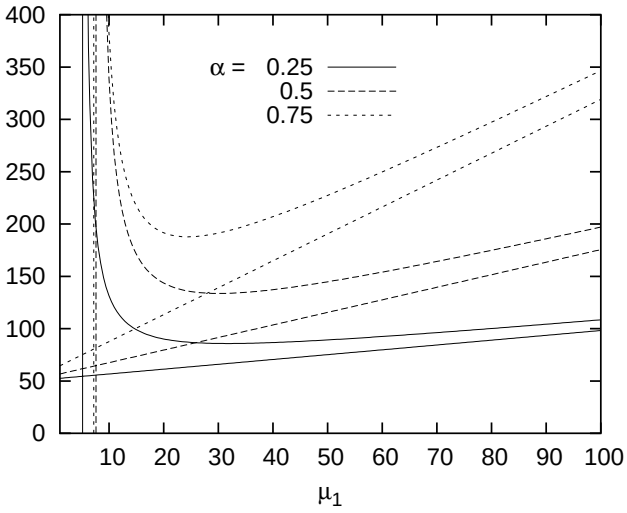
In Figuur S.20 tonen we de gemiddelde klasse-2 vertragingstijd als functie van de totale aankomstbelasting voor constante bedieningstijden van 2 slots voor de klasse-1 pakketten, voor constante bedieningstijden van 20 slots voor de klasse-2 pakketten, voor $\alpha = 0.25, 0.5$ en 0.75 en voor de PR, PRD en PRI prioriteitsdisciplines. Verder hebben we ook de verticale asymptoten van de PRD en PRI curven getoond. Deze tonen voor welke aankomstbelasting de effectieve belasting gelijk wordt aan 1. Deze twee soorten belastingen zijn niet identiek aangezien de herhalingen aanleiding geven tot 'extra' belasting. Rechts van deze asymptoten verliezen de PRI en PRD wachtlijnen dus hun stabiliteit. Merk op dat voor deterministische klasse-2 bedieningstijden de PRD en PRI disciplines aanleiding geven tot dezelfde discipline (aangezien bij herhaling van de klasse-2 bedieningstijd een nieuw monster nemen en de oorspronkelijke bedieningstijd behouden identiek zijn). De curven in Figuur S.20 zijn dan ook identiek voor deze 2 disciplines. De figuur toont dat de gemiddelde klasse-2 bedieningstijd significant groter kan zijn in het geval van PRI en PRD t.o.v. het PR geval. Dit is een gevolg van de extra belasting die door de herhalingen van de klasse-2 bedieningen wordt toegevoegd.



Figuur S.21: Gemiddelde klasse-2 vertragingstijd versus de totale belasting voor de PRD (onderste curven) en PRI (bovenste curven) prioriteitsdisciplines voor variabele bedieningstijden ($\mu_1 = \mu_2 = 20$)

Om de invloeden van de PRD en PRI prioriteitsdisciplines onderling te vergelijken veronderstellen we in de volgende figuren de klasse-2 bedieningstijden variabel. In Figuur S.21 is de gemiddelde klasse-2 vertragingstijd uitgezet voor de PRD en PRI prioriteitsdisciplines als functie van de totale aankomstbelasting voor constante bedieningstijden van 20 slots voor de klasse-1 pakketten. De klasse-2 bedieningstijden zijn gelijk aan 10 slots of 30 slots elk met kans 1/2. Verder is $\alpha = 0.25, 0.5$ en 0.75. Het is duidelijk dat de PRI prioriteitsdiscipline aanleiding geeft tot (gemiddeld) langere vertragingstijden. Dit komt doordat in de PRD prioriteitsdiscipline lange klasse-2 bedieningstijden (30 slots) herbemonsterd kunnen worden tot korte bedieningstijden (10 slots) bij onderbreking door klasse-1 pakketten. Het omgekeerde kan natuurlijk ook maar de kans dat een lange bedieningstijd onderbroken wordt is groter dan dat een korte onderbroken wordt. Daardoor is de 'netto' invloed van de herbemonstering op de prestatie van het systeem positief.

Figuur S.22 tenslotte toont de gemiddelde klasse-2 vertragingstijd in het geval van de PR, PRD en PRI prioriteitsdisciplines als functie van de gemiddelde klasse-1 bedieningstijden. De klasse-1 en klasse-2 bedieningstijden zijn constant met $\mu_2 = 20$ slots, $\rho_T = 0.75$ en $\alpha = 0.25, 0.5$ en 0.75. Opnieuw zijn de PRD en PRI disciplines identiek wegens de constante klasse-2 bedieningstijden. In het geval van de PR discipline groeit de gemiddelde klasse-2 vertragingstijd met stijgende μ_1. In de PRD en PRI gevallen zijn er twee tegenwerkende effecten: enerzijds geven langere klasse-1 pakketten aanleiding tot langere ophopingsperiodes voor de klasse-2 pakketten in de buffer, waar-



Figuur S.22: Gemiddelde klasse-2 vertragingstijd versus de klasse-1 bedieningstijden voor de PR (onderste curven) en PRD en PRI (bovenste curven) prioriteitsdisciplines voor constante bedieningstijden ($\rho_T = 0.75, \mu_2 = 20$)

door de gemiddelde klasse-2 vertragingstijd stijgt (dit is ook het effect dat in het PR geval speelt). Anderzijds betekenen langere klasse-1 pakketten minder klasse-1 pakketten (aangezien we de (klasse-1) belasting constant houden) waardoor de kans dat een klasse-2 bediening onderbroken wordt door nieuw aankomende klasse-1 pakketten kleiner wordt. Dit heeft een dalend effect op de gemiddelde klasse-2 vertragingstijd. Op de figuur kan gezien worden dat - voor kleine klasse-1 bedieningstijden - de gemiddelde vertragingstijd spectaculair stijgt naarmate de pakketten korter worden. Dit komt wegens het tweede effect: als μ_1 klein is, komen heel veel korte klasse-1 pakketten aan in het systeem waardoor de klasse-2 bedieningen veel onderbroken worden. Aangezien deze steeds herhaald moeten worden na een onderbreking heeft dit aanleiding tot heel grote vertragingstijden. Voor hoge μ_1 stijgt de gemiddelde klasse-2 vertragingstijd met stijgende μ_1 en dat is een gevolg van het eerste effect. Het is ook duidelijk dat voor lange klasse-1 pakketten het verschil tussen de PRD en PRI disciplines enerzijds en de PR discipline anderzijds relatief klein is. Aangezien de klasse-1 pakketten lang zijn, komen er weinig aan. Hierdoor worden onderbrekingen van klasse-2 bedieningen schaars en is ook de invloed van voortzetten of herhalen van de bediening na een onderbreking minder belangrijk. Door deze twee tegenwerkende effecten krijgen we een optimum voor μ_1 in de PRD en PRI prioriteitwachtlijnen waarvoor de gemiddelde klasse-2 vertragingstijd minimaal wordt.

S.4 Conclusies

In dit proefschrift hebben we een gedetailleerde studie beschreven van *discrete-tijd* wachtlijnmodellen met verschillende types *prioriteitsdisciplines*. Doorheen het proefschrift hebben we twee prioriteitsklassen beschouwd. Het aankomstproces is i.i.d. van slot-tot-slot, alhoewel de aantallen aankomsten in één slot van beide klassen gecorreleerd kunnen zijn. In een eerste model hebben we ons beperkt tot constante bedieningstijden van één slot en dit is in verdere modellen uitgebreid naar willekeurige bedieningstijden. Verder hebben we in deze latere modellen geïncorporeerd dat de bedieningstijden van verschillende prioriteitsklassen verschillende distributies kunnen hebben.

We hebben dus eerst een prioriteitswachtlijn met constante bedieningstijden van één slot bestudeerd (in hoofdstuk 2 van het Engelstalig gedeelte). Dit is een vrij eenvoudig model en geeft aanleiding tot een gesimplificeerde analyse (t.o.v. de latere analyses) zodat de lezer voeling krijgt met het gebruik van pgf's in de analyse van wachtlijnen met prioriteiten en de manier waarop de prestatiematen uit deze pgf's kunnen gehaald worden. Maar dit (simplistisch) model is ook bruikbaar in de praktijk, b.v., in telecommunicatienetwerken waar de pakketten die door het netwerk getransporteerd worden alle van dezelfde lengte zijn (b.v. ATM).

Vervolgens is dit model uitgebreid naar algemene bedieningstijden. Een aantal verschillende prioriteitsdisciplines zijn beschreven en geanalyseerd. Ten eerste de niet-preëmptieve prioriteitswachtlijn. In deze wachtlijn worden bedieningen van pakketten niet onderbroken. De analyse van dit type wachtlijnen is in hoofdstuk 3 van het Engelstalig gedeelte van dit proefschrift beschreven. Vervolgens hebben we in hoofdstuk 4 van het Engelstalig gedeelte de preëmptieve prioriteitswachtlijn met voortzetting geanalyseerd. Bij deze discipline wordt de bediening van een lage-prioriteitspakket onderbroken door nieuw aankomende hoge-prioriteitspakketten. De onderbroken bediening wordt later voortgezet. Uiteindelijk hebben we in hoofdstuk 5 twee preëmptieve prioriteitswachtlijnen met herhaling bestudeerd. Het verschil met de vorige prioriteitsdiscipline is dat de onderbroken bediening in dit laatste hoofdstuk herhaald wordt vanaf het begin. Het pakket moet dus opnieuw volledig bediend worden.

Doorheen de verschillende bestudeerde wachtlijnmodellen, hebben we een vrij algemene analytische methode gebruikt. Eerst hebben we een Markov keten voor het desbetreffende systeem geconstrueerd. Vervolgens hebben we een methode gebaseerd op probabiliteitsgenererende functies gebruikt om de gezamenlijke pgf van de stochastische veranderlijken gedefinieerd in de Markov keten (in regime) te berekenen. Startende van deze gezamenlijke pgf worden dan alle verdere pgf's die onze interesse wegdragen afgeleid. Uit deze pgf's verkrijgen we tenslotte de nuttige prestatiematen, zoals de momenten en staartprobabiliteiten van verschillende toevalsveranderlijken.

Alhoewel we dit proefschrift als een afgewerkt geheel hebben voorgesteld, is

onderzoek natuurlijk nooit voltooid. Er zijn dan ook nog velerlei mogelijke uitbreidingen van de modellen en analyses gepresenteerd in dit proefschrift. Zo kunnen de modellen uitgebreid worden naar meer dan twee prioriteitsklassen, tijdscorrelatie in het aankomstproces, prioriteitssystemen met meerdere bedieningsstations, analyse van het uitgangsproces van prioriteitswachtlijnen, transiëntanalyses, ...

Chapter 1

Introduction

1.1 Queues and buffers

Queues are part of daily life. Queueing phenomena are observed on the roads, when waiting in line in supermarkets, movie theaters, post offices, emergency rooms in hospitals, when being on a waiting list for surgery, ... In general a *queueing process* can be defined as the process of waiting before getting some kind of service.

Specifically in telecommunication networks, buffers are used to store information that cannot be sent instantly to its next destination. The cause of this is the instantaneous overload of arriving information, i.e., during a time period more information arrives than can be (simultaneously) transmitted. Examples of causes are multiplexing of several input links/traffic streams to one output link, switching from a link to a slower link, temporary errors in output links, re-sequencing of information units, ... Without buffers, too much information would get lost in the above described cases.

The entities that arrive in the queue are in general called *units* throughout this dissertation. We will however also use - the telecommunications inspired - *cells* or *packets*.

1.2 Importance of the queue/buffer behavior

The behavior of a queue (in time) is an important topic, e.g., when designing new roads it is beneficial to know beforehand whether a certain design is prone to the formation of queues.

In case of a waiting list for surgery, people with life-threatening injuries may be part of the queue. It is thus important to study the behavior of the queue -

or more specifically the amount of time a patient is on the waiting list before surgery - in order to avoid unnecessary casualties.

In telecommunications, the way buffers behave is an important research topic because the performance of the network and the Quality of Service (QoS) experienced by the users is closely related to the buffers' behavior. Information can be lost because of buffer overflow or information units can suffer too long delays. The consequences to the users will vary depending on the application. For instance, large delays (and delay variations) are very bad for real-time applications (voice, video, ...), while they are more acceptable for non-real-time applications (data). On the other hand, loss of information is devastating for data, while it is to some extent acceptable for voice (because of redundancy). Therefore, the consequences for the experienced QoS depend on the application, but it is clear that the buffer behavior plays a major role in (how the user evaluates) the performance of the network.

1.3 Stochastic variables

In order to study the queue behavior, first a number of input and output *variables* have to be defined. The *input* variables describe the characteristics of the traffic offered to the queue. The *output* variables describe the queue behavior. Since traffic behavior is of a non-deterministic, probabilistic nature - and thus also the queue behavior is of an uncertain nature - these variables are defined as *stochastic* variables. In this dissertation, the input variables will be assumed to be known variables, while the output variables are the variables which have to be analyzed. The input stochastic variables used in this dissertation will be discussed in more detail in section 1.7. The output stochastic variables, i.e., the variables that describe the behavior of the particular queues that will be analyzed in this dissertation, will be discussed in section 1.8.

1.4 Analysis techniques

There are several analysis techniques, all with their own specific advantages and disadvantages. They can roughly be categorized in 4 groups. The first technique is the *analytical* technique. System equations for the stochastic variables of interest are established and solved analytically. The second way of analyzing queueing systems is through a *numerical* method. In this approach, the system equations of the desired stochastic variables are also determined, but they are solved numerically. The third method is solving the system through *simulations*. In this method, a computer program is written to simulate the network/queueing system and results are calculated by running the program. The last approach is the *experimental* approach. Instead of a

computer-version of the (queueing) system the 'real thing' is built and experiments are executed on this. The variables of interest are measured while running the experiments. In telecommunications, e.g., one can build a test-bed version of the network to perform experiments on.

As already mentioned, each method has its own advantages and disadvantages. The advantage of the analytical method is the obvious parameter-dependence of the results. In general, formulas are obtained in which the different parameters appear. Changing the value of a parameter gives the (new) result immediately. The disadvantage of the analytical method is the need to use simplified models in order to be able to analyze the system under consideration. The important question is whether the simplified model is still "rich" enough to model the real system in a satisfying way. The advantages and disadvantages of the experimental method are the opposite of the analytical method. Its advantage is that the real system under consideration is built and the obtained results are thus certainly reliable. The disadvantage is that parameter-dependence is totally lacking. If a parameter has to be changed, the experiment has to be repeated. Therefore, this method takes considerably more time when the influence of certain parameters has to be studied. Furthermore, this method is not always practically possible (e.g., when designing roads) or too expensive. The numerical analysis and simulation approaches lay between those two extremes, where the advantages and disadvantages of the numerical analysis lean more to analytical analysis while those of the simulations approach lean more to the experimental approach.

In this dissertation, we use an analytical technique based on *probability generating functions* (see further for more details). We will thus propose a mathematical model and use this analytical technique to study this model. In order to check some results which are only found through the use of an approximative solution technique, we will compare with simulations in order to test the validity of the approximation.

1.5 Multiple types of traffic

In many queueing studies, traffic is assumed to be homogeneous, i.e., all traffic is assumed to be of the same type. However, most traffic is heterogeneous in nature. It can thus be subdivided in multiple classes, depending on both their characteristics and requirements. Indeed, the traffic characteristics are diverse (bursty traffic $\leftrightarrow$ monotonous traffic, long service times or short service times, constant $\leftrightarrow$ variable service times, ...). Secondly, different traffic types can have different requirements, e.g., different loss requirements (i.e., the probability of not receiving any service and "being lost" in telecommunication networks), different delay requirements, ...

An example of a classification, which is frequently used in nowadays multimedia telecommunication systems, is *real-time* traffic (e.g., voice) and *non-real-*

time traffic (e.g., data). A similar classification can be used in case of a waiting list for surgery: some patients' treatment is extremely urgent, while other patient can sustain some time (days, weeks, ...) before receiving treatment. This classification is the most obvious - in case of these two examples - but other and/or more detailed classifications are possible.

1.6 Scheduling multiple types of traffic

As touched upon in the previous section, it is sometimes necessary to differentiate traffic classes because of different requirements. In telecommunications, different types of traffic need different QoS standards. For real-time applications, it is important that mean delay and delay-jitter are bound, while for non-real-time applications, the loss ratio is the restrictive quantity.

In the remainder of this section, we will give a brief overview of scheduling schemes that have been proposed - and analyzed - in order to guarantee acceptable delay bounds to delay-sensitive traffic in multimedia networks. Amongst these scheduling disciplines are some well-known strategies like weighted round-robin (WRR), weighted fair queueing (WFQ) or generalized processor sharing (GPS), earliest deadline first (EDF), probabilistic priority (PP) and (strict) priority scheduling.

In a queueing system with WRR (see e.g. Liu et al. [1997] and references therein), WFQ or GPS (see e.g. Parekh and Gallager [1994] and references), the server serves a number of queues by a weighted schedule. Delay sensitive traffic is assigned a higher weight, i.e., (on average) delay-sensitive traffic is served earlier/longer than delay-insensitive traffic.

When the EDF scheduling is applied, deadlines are imposed on the packets that have to be served (based on their QoS constraints) and packets are transmitted in the order of their deadlines (see Liebeherr and Wrege [1999] and references therein).

A PP scheduling discipline (see e.g. Tham et al. [2002]) serves a given number of priority queues in a probabilistic manner. Each priority queue is assigned a parameter p_i, which determines the probability that a packet from that priority queue is served when the server is ready to transmit a (new) packet.

All these scheduling disciplines try to give some kind of priority to delay-sensitive traffic over delay-insensitive traffic. The most drastic in this respect is the strict priority scheduling (which we analyze in this dissertation). With this scheduling, as long as delay-sensitive (or high-priority) packets are present in the queueing system, this type of traffic is served. Delay-insensitive packets can thus only be transmitted when no delay-sensitive traffic is present in the system. As already mentioned, this is the most drastic way to meet the QoS constraints of delay-sensitive traffic (and thus the scheduling with the most disadvantageous consequences on the delay characteristics of the delay-insensitive traffic), but also the easiest to implement. E.g., in Shenker [1995]

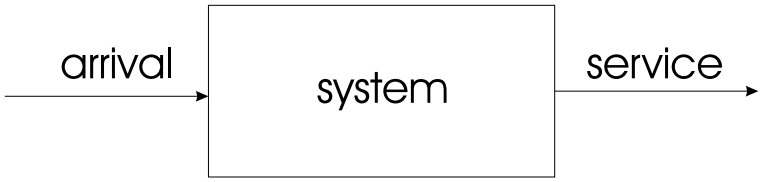


Figure 1.1: Conceptual representation of a queueing system

categorizing traffic in different service classes and using a priority scheduling discipline in the routers to serve these classes are proposed as the best solution in the nowadays and future Internet.

Note that the application of priority scheduling is not limited to telecommunications. Basically all queues in which the delay requirements of the units differ may profit from a priority scheduling discipline.

A last notable example is a waiting list for surgery. According to Wang [2004], a story entitled "Surgeon can't sleep as patients die waiting" appeared in the Toronto Star newspaper on May 13, 1988. It is about a senior cardiovascular surgeon who announced that there had been twelve or thirteen deaths on his personal waiting list for open-heart surgery. One of the reasons was that patients (and their diseases) are not carefully evaluated and patients were picked for surgery more or less on a First-Come-First-Served basis. This triggered a change in the scheduling of the patients for surgery (see Hadorn [2000], Wang [2004]): (the diseases of) the patients are evaluated and are put in a certain priority class according to the life-threatening nature of the disease. The patients are scheduled for surgery according to their priority.

1.7 Queueing model

In this section, we describe the characteristics of the queueing model under investigation in the remainder of this dissertation. *Units* arrive in the system, wait a certain amount of time before starting to receive service and when their service is completed, they leave the system. The modeling assumptions can roughly be split in three parts (see Figure 1.1). Firstly, we specify the *buffer system* itself. Secondly, the *arrival process* is modeled and finally the *service process* is characterized. We will put special emphasis on the characterization and impact of the priority scheduling discipline, studied in this dissertation.

1.7.1 System modeling

We consider a *discrete-time single-server multi-class* queueing system with infinite buffer space and a priority scheduling discipline. Time is divided into slots

of equal length. There is one server which transmits the units from the queue. Units which cannot be transmitted instantaneously are stored in the queue. Since we analyze discrete-time queueing systems, *all* stochastic variables encountered in this dissertation are assumed to only take discrete (non-negative) values. The buffer space is assumed to be infinitely large, which means that no units are lost. Service of units can only start at slot boundaries. Notice that this means that an arriving unit cannot enter the server during its arrival slot, even when the server is empty when the unit arrives. Traffic is generally subdivided into M different classes, but we only discuss and analyze systems with *two* priority classes in this dissertation, i.e., $M = 2$ throughout the dissertation. This is done for a couple of reasons. The first is the fact that most of the time two (priority) classes are sufficient, e.g. in telecommunications systems, one class represents the real-time traffic while the non-real-time traffic is categorized by a second class. Secondly, although the techniques used for a system with 2 classes are (more or less straightforwardly) extendable to a system with a general number of classes, the formulas become more cumbersome and give little extra information.

The priority scheduling discipline of the buffer will be thoroughly discussed in subsection 1.7.4.

1.7.2 Arrival process

In most queueing studies, especially in continuous-time queueing analyses, the arrival process is characterized by the *interarrival time*. This stochastic variable is defined as the time between two consecutive arrival epochs of units. In discrete-time queues, an alternative characterization can be used. In this characterization, the *number of per-slot arrivals* is characterized. This equals the number of units that arrive in a (random) slot. We use the latter characterization of the arrival process throughout this dissertation. Since we only focus on slot boundaries and on discrete stochastic variables (see further) the precise moment a unit arrives in the slot is of no importance in our analysis, but this can be altered if one does not want to restrict the analysis to slot boundaries and/or discrete stochastic variables (see e.g. Bruneel [1993] for more details).

Units of two types of traffic arrive in the system, namely units of class-1 and of class-2. We denote the number of arrivals of class-j during slot k by $a_{j,k}$ ($j = 1, 2$). Both types of unit arrivals are assumed to be independent and identically distributed (i.i.d.) from slot-to-slot and are characterized by the joint probability mass function (pmf)

$$a(n_1, n_2) \triangleq \text{Prob}[a_{1,k} = n_1, a_{2,k} = n_2], \tag{1.1}$$

and corresponding joint probability generating function (pgf) $A(z_1, z_2)$,

$$A(z_1, z_2) \triangleq \text{E}\left[z_1^{a_{1,k}} z_2^{a_{2,k}}\right] \tag{1.2}$$

$$= \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} a(n_1, n_2) z_1^{n_1} z_2^{n_2}. \tag{1.3}$$

We furthermore denote the total number of arriving packets during slot k by $a_{T,k} \triangleq a_{1,k} + a_{2,k}$ and its pgf is defined as

$$A_T(z) \triangleq \mathrm{E}[z^{a_{T,k}}] \tag{1.4}$$
$$= A(z, z). \tag{1.5}$$

In the same way, we define the marginal pgf of the number of arrivals from class-j during a slot by

$$A_j(z) \triangleq \mathrm{E}[z^{a_{j,k}}] \tag{1.6}$$
$$= A(z_1, z_2)\big|_{z_j=z, z_i=1, i\neq j}, \tag{1.7}$$

$j = 1, 2$. We furthermore denote the (mean) arrival rate of class-j units by $\lambda_j = A_j'(1)$ and the total arrival rate by $\lambda_T = A_T'(1) = \lambda_1 + \lambda_2$.

Notice that equations (1.1) and (1.3) incorporate the possibility that the numbers of arrivals from different classes (within a slot) are correlated. This is sometimes called a *discrete structured batch arrival process* or *structured input*. Incorporating this correlation is for instance necessary in the case of a non-blocking output-queueing switch with N inlets and N outlets (see Figure 1.2). The numbers of arrivals on each inlet are assumed to be i.i.d., and generated by a Bernoulli process with arrival rate λ_T. An arriving unit is assumed to be of class-j with probability λ_j/λ_T, $j = 1, 2$ (with $\lambda_1 + \lambda_2 = \lambda_T$). The incoming units are then routed to the output queue corresponding to their destination in an independent and uniform way. Therefore, the output queues behave identically and we can concentrate on the analysis of one output queue. The former arrival process assumptions lead to the fact that the arrivals of both types of units to an output queue are generated according to a two-dimensional binomial process. It is fully characterized by the following joint pgf

$$A(z_1, z_2) = \left(1 + \sum_{j=1}^{2} \frac{\lambda_j}{N}(z_j - 1)\right)^N. \tag{1.8}$$

Obviously, the numbers of class-j arrivals at an output queue during a slot are mutually correlated (for finite N). This is simply demonstrated by the following observation: when m class-i units arrive at the tagged queue during a slot ($0 \leq m \leq N$), the maximum number of arrivals of the other class during the same slot is limited by $N - m$. We note that for N going to infinity, expression

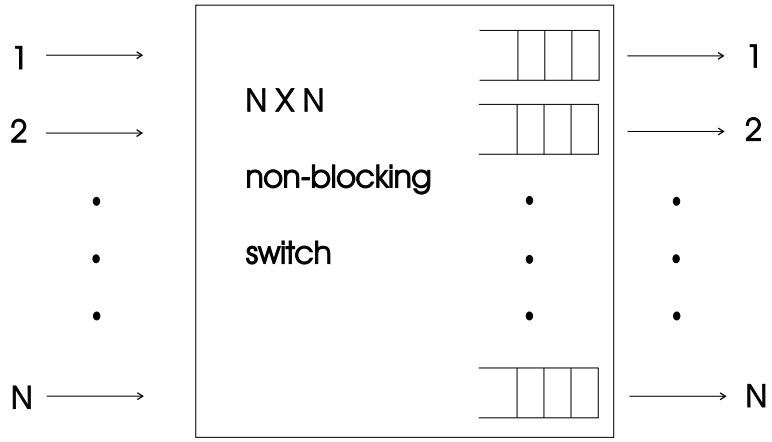


Figure 1.2: An NxN output queueing switch

(1.8) becomes

$$A(z_1, z_2) = e^{\sum_{j=1}^{2} \lambda_j (z_j - 1)} \tag{1.9}$$

$$= \prod_{j=1}^{2} e^{\lambda_j (z_j - 1)}, \tag{1.10}$$

a product of two pgf's of Poisson distributions with means λ_j ($j = 1, 2$) respectively, and as a result, the numbers of arrivals of both classes are mutually uncorrelated. We use this type of switch extensively as a means to show the applications of our results. The applications are however not limited to this type of telecommunication device - and not even to telecommunications.

1.7.3 Service process

The service times of consecutive units are assumed to be independent. Furthermore, the service times of the consecutive class-j units, $j = 1, 2$, are assumed to be i.i.d. Therefore, we define the stochastic variable s_j as the service time of a random class-j unit (expressed in slots). s_j is characterized by its pmf

$$s_j(n) \triangleq \text{Prob}[s_j = n \text{ slots}],\ n \geq 1, \tag{1.11}$$

and pgf

$$S_j(z) \triangleq \text{E}[z^{s_j}] \tag{1.12}$$

$$= \sum_{n=1}^{\infty} s_j(n) z^n, \tag{1.13}$$

with $j = 1, 2$. Note that serving a unit requires at least one slot. We furthermore denote the mean service time of a class-j packet by $\mu_j \triangleq \mathrm{E}[s_j] = S_j'(1)$. We define the arrival load offered by class-j packets as $\rho_j \triangleq \lambda_j \mu_j$ ($j = 1, 2$). The total load is then given by $\rho_T \triangleq \rho_1 + \rho_2$.

In this dissertation, a number of specific types of distributions for the service times are important. Firstly, in chapter 2, we will consider packets with deterministic service times of 1 slot. Secondly, the shifted geometric distribution will play an important role. The pmf $x(n)$ and pgf $X(z)$ of a random variable X with a shifted geometric distribution with parameter β are given by

$$x(n) = (1 - \beta)\beta^{n-1}, \; n \geq 1, \tag{1.14}$$

and

$$X(z) = \frac{(1 - \beta)z}{1 - \beta z}, \tag{1.15}$$

respectively. The reason for using this specific distribution for the service times as a first attempt to analyze a specific queueing system, is its memoryless property. In our analysis, this property implies that at a certain time instant the remaining number of slots that the unit in service still has to receive service is independent of the number of slots it is already in the server. In the corresponding analysis, we do not have to keep track of the latter variable, as opposed to in a system with (more) general service times.

1.7.4 Priority scheduling discipline

Class-1 units are assumed to have priority over class-2 units and within one class the service discipline is First Come First Served (FCFS). Due to the priority scheduling mechanism, it is as if class-1 units are stored in front of class-2 units in the queue. So, if there are class-1 packets in the queue when the server is ready to service a new unit, the class-1 unit with the longest waiting time will start service. Only if there are no class-1 units, a class-2 unit - namely the one with the longest waiting time - starts service.

In the case that the service time may be more than one slot, two types of priority scheduling are distinguished, namely, *non-preemptive* and *preemptive* priority scheduling. In the non-preemptive priority scheduling discipline, service of a unit cannot be interrupted. So, when a unit of a certain class is being served its service is not interrupted by newly arriving units, even if they have higher priority. The latter units have to wait until the lower priority unit is totally served. In the preemptive priority scheduling discipline

on the other hand, the service of that unit will be interrupted by newly arriving higher priority traffic and the unit whose service time was interrupted will have to go back to the queue. In the latter discipline, we can characterize three categories, namely, the preemptive *resume*, the preemptive *repeat different* (or *repeat with resampling*) and the preemptive *repeat identical* (or *repeat without resampling*) priority scheduling disciplines. They differ from each other in the way they handle a unit whose service was interrupted, at the moment it enters the server for a second (or third, fourth,...) time. In the preemptive resume priority scheduling discipline, the unit resumes service where it was interrupted, i.e., only the part which was not yet served before the interruption, has to be served afterwards. In the preemptive repeat priority scheduling disciplines, the whole unit has to be served after the interruption by higher priority traffic, thus including the part that was already served before. The preemptive repeat different priority discipline takes a new sample with the same distribution. The service time of that particular packet may thus change after an interruption. The preemptive repeat identical on the other hand keeps the same service time when re-attempting to transmit the unit.

In the remainder of this dissertation, we will use the following abbreviations for the different priority scheduling disciplines: *NP* for the non-preemptive, *PR* for the preemptive resume, *PRD* for the preemptive repeat different and *PRI* for the preemptive repeat identical priority scheduling discipline.

1.8 Typical results

1.8.1 Output stochastic variables

As discussed in section 1.3, a number of output stochastic variables can be defined and have to be analyzed. A first important stochastic variable is the *system contents* at the beginning of a random slot. The system contents is defined as the number of units in the system. In the two-class priority queues we analyze in this dissertation, the system contents of class-j at the beginning of slot k is denoted by $u_{j,k}$, and equals the number of units of class-j in the system at the beginning of slot k. The total number of units in the system - or the total system contents - at the beginning of the k-th slot is denoted by $u_{T,k}$ and is given by $u_{1,k} + u_{2,k}$.

A related stochastic variable is the *queue contents*, which is defined as the number of units in the queue at a certain time instant. The difference between the system contents and the queue contents is that the possible unit being served at that time instant is not included in the latter. In the remaining chapters of this dissertation, the queue contents of class-j at the beginning of slot k will be denoted by $q_{j,k}$ and the total queue contents by $q_{T,k} = q_{1,k} + q_{2,k}$.

A third important characteristic is the *unfinished work* at a given time instant, which is defined as the number of slots it takes to empty the system of all

units stored at that time, if there are no new arrivals (from that time instant onwards). The unfinished work of class-j is defined as the number of slots it would take to serve all class-j units stored at that time, if there were no new arrivals and if the server only served the class-j packets from that time instant onwards. The unfinished work of class-j at the beginning of slot k is denoted by $w_{j,k}$ and

$$w_{T,k} = w_{1,k} + w_{2,k}, \tag{1.16}$$

is the total unfinished work at the beginning of the k-th slot. From expression (1.16) an alternative explanation of the unfinished work of a single class (class-j) can be given: this unfinished work is the number of slots of the *total* unfinished work that the server spends on serving class-j units.

In the analyses in this dissertation, we will concentrate on the steady-state versions of these variables, i.e., we will find performance measures of the stochastic variables defined so far for $k \to \infty$. In the remainder of this subsection we discuss some more stochastic variables, but we directly define the steady-state versions of these stochastic variables (in contrast with the system contents, queue contents and unfinished work).

The steady-state *delay* of a specific unit is defined as the time a unit spends in the system, i.e., the time period between its arrival and departure instants. As discussed before, in discrete-time queueing systems, it is common practice to only consider the discrete-part of the delay. Or, more precisely, the delay of a specific unit is defined as the number of slots between the end of the arrival slot of this unit and the end of its departure slot. In the remaining chapters of this dissertation, the steady-state delay of a class-j unit is denoted by d_j, while the steady-state delay of a random unit is denoted by d.

There is a related stochastic variable to the delay, namely the *queueing* or *waiting time*, which is defined as the amount of time the unit stays in the queue before starting service. t_j is defined as the (steady-state) waiting time of a class-j unit and t is the waiting time of a random unit.

1.8.2 Performance measures

Probability generating functions

For the stochastic variables defined in 1.8.1, several specific performance characteristics can be defined and calculated. Since we make extensive use of probability generating functions in this dissertation, pgf's of the stochastic variables of interest will be calculated first. For the system contents, queue contents and unfinished work, the *joint* pgf of these (steady-state) stochastic variables of all classes will be calculated. If X_j, $j = 1, 2$ is the steady-state stochastic variable (of interest) of class-j, then the joint pgf of the stochastic

variables X_1 and X_2 is defined as

$$X(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{X_1} z_2^{X_2}\right]. \tag{1.17}$$

From this joint pgf, the marginal pgf's can be calculated as follows

$$\begin{aligned} X_j(z) \triangleq & \mathrm{E}\left[z^{X_j}\right] && (1.18)\\ = & X(z_1, z_2)\Big|_{z_j=z, z_i=1, i\neq j}. && (1.19) \end{aligned}$$

Also, the sum of the stochastic variables in question can be found from (1.17), yielding

$$\begin{aligned} X_T(z) \triangleq & \mathrm{E}\left[z^{X_T}\right] && (1.20)\\ = & \mathrm{E}\left[z^{X_1+X_2}\right] && (1.21)\\ = & X(z, z). && (1.22) \end{aligned}$$

For the two other types of stochastic variables, notably the steady-state delay and waiting time, the pgf's $X_j(z)$ of these stochastic variables of class-j, $j = 1, 2$ and the pgf $X(z)$ of the stochastic variable of a random unit will all be calculated separately.

Moments

From the obtained pgf's, the (central) moments of the variables can be calculated. For instance, the *mean value* and *variance* of a stochastic variable X with pgf $X(z)$ are given by

$$\begin{aligned} \mathrm{E}[X] = & X'(1) && (1.23)\\ \mathrm{Var}[X] \triangleq & \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] && (1.24)\\ = & X''(1) + X'(1) - (X'(1))^2. && (1.25) \end{aligned}$$

So, by calculating the derivatives of the pgf's of the stochastic variables one finds the moments of these stochastic variables. Note that this is not restricted to the calculation of the mean value and the variance, but that, in principle, all (central) moments of a stochastic variable can be derived once its pgf is obtained. However, in order to obtain the n-th (central) moment up to the n-th derivative of the pgf has to be calculated, which makes the calculation of high moments practically infeasible. In this dissertation, we will restrict our results to mean values and variances.

For the stochastic variables of which the joint pgf is obtained (see expression (1.17)), the concept of moments of the respective stochastic variables separately can be expanded to cross-moments of the respective stochastic variables. For instance, the *covariance* of two stochastic variables X_1 and X_2 is given by

$$\begin{aligned} \text{Cov}[X_1, X_2] \triangleq & \text{E}\left[(X_1 - \text{E}[X_1])(X_2 - \text{E}[X_2])\right] && (1.26) \\ = & \left.\frac{\partial^2 X(z_1, z_2)}{\partial z_1 \partial z_2}\right|_{z_1=z_2=1} - X_1'(1)X_2'(1), && (1.27) \end{aligned}$$

with $X(z_1, z_2)$ the joint pgf of X_1 and X_2 and $X_j(z)$ the (marginal) pgf of X_j ($j = 1, 2$). The *correlation coefficient* is the normalized covariance and is defined as

$$\text{Corr}[X_1, X_2] \triangleq \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var}[X_1]\text{Var}[X_2]}}, \tag{1.28}$$

and is a real number between -1 and $+1$.

Probability mass functions and tail probabilities

Another important performance characteristic is the *complete distribution* or the *probability mass function* of the stochastic variables. The distribution of a (discrete) stochastic variable X is defined as

$$x(n) = \text{Prob}[X = n], \tag{1.29}$$

for all non-negative discrete n. This pmf can, in principle, be straightforwardly derived from the corresponding pgf of the stochastic variable by means of the so-called probability generating property of probability generating functions, yielding

$$x(n) = \frac{1}{n!}\left.\frac{d^n X(z)}{dz^n}\right|_{z=0}. \tag{1.30}$$

Since the n-th derivative of the pgf $X(z)$ is required for the calculation of $x(n)$, this is in general not a practical method, especially for high n. Since we are most of the time (only) interested in $x(n)$ for high n, the so-called *tail probabilities* of X, a more practical method has to be used. A possible solution is the numerical inversion of pgf's (using e.g. Discrete Fourier Transforms). In this dissertation, we will use an approximate analytical technique which is introduced in the remainder.

Quite some research has been done on finding (good) analytic approximations - which are easy to calculate - of the $x(n)$ once (some information of) $X(z)$ is known. This is not only a research topic in queueing theory (see e.g.

Bruneel et al. [1994], Laevens [1999] and others), but also in combinatorics (see e.g. Bender [1974], Flajolet and Odlyzko [1990] and references therein). Basically, to calculate the complete distribution of a stochastic variable, we have to find/know all singularities of its pgf, or more precisely of its analytic continuation (for more details see Bruneel et al. [1994]). If we are only interested in the tail probabilities on the other hand, it suffices to determine the dominant singularities of the pgf, i.e., the singularities with smallest modulus. These are nothing but the singularities on the circle of convergence of the pgf. We will frequently make use of some basic (transfer) theorems in this dissertation. Before we describe these theorems we first give some notations: $F(z) \sim G(z)$ means that $F(z)/G(z) \to 1$ as z goes to a certain value (which should be clear from the context). Similarly, $f(n) \sim g(n)$ means that $f(n)/g(n) \to 1$ as $n \to \infty$. Finally, $f(n) = o(g(n))$ means that $f(n)/g(n) \to 0$ as $n \to \infty$.

The first theorem is Darboux's theorem:

Theorem 1.1 (Darboux's theorem) *Suppose $X(z) = \sum_{n=0}^{\infty} x(n)z^n$ with positive real coefficients $x(n)$ is analytic near 0 and has only algebraic singularities α_k on its circle of convergence $|z| = R$, in other words, in a neighborhood of α_k we have*

$$X(z) \sim (1 - \frac{z}{\alpha_k})^{-\omega_k} G_k(z), \tag{1.31}$$

where $\omega_k \neq 0, -1, -2, \ldots$ and $G_k(z)$ denotes a nonzero analytic function near α_k. Let $\omega = max_k Re(\omega_k)$ denote the maximum of the real parts of the ω_k. Then we have

$$x(n) = \sum_j \frac{G_j(\alpha_j)}{\Gamma(\omega_j)} n^{\omega_j - 1} \alpha_j^{-n} + o(n^{\omega - 1} R^{-n}), \tag{1.32}$$

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

So, once a pgf is *explicitly* calculated, it suffices to find all algebraic singularities on the circle of convergence and use Darboux's theorem. Following theorem (see e.g. Laevens [1999]) gives one of those singularities:

Theorem 1.2 (Vivanti's theorem) *If $X(z)$ is a power series with real positive coefficients and with radius of convergence R, then $z = R$ is a singularity of $X(z)$.*

This theorem is sometimes also attributed to the German mathematician Pringstein.

The point of the circle of convergence on the real positive axis is thus always a singularity of the corresponding pgf. In the remainder, we assume that this singularity is the only singularity on the radius of convergence. Extensions can be made to cases where this assumption is not valid (some extensions are not straightforward though), but since this assumption seems to be valid for

a whole range of "typical" distributions (or in other words, the underlying distributions have to be of a very specific nature for the assumptions *not* to be valid), we will not go into these cases. We then "translate" Darboux's theorem in this case as follows: if $X(z)$ is explicitly known and in the neighborhood of its *dominant* singularity R on the positive real axis we have

$$X(z) \approx \frac{G(z)}{(R-z)^{\omega}}, \tag{1.33}$$

then

$$x(n) \approx \frac{K_X}{\Gamma(\omega)} n^{\omega-1} R^{-n-\omega}, \tag{1.34}$$

for n large enough and with $K_X = G(R)$. For example, when the dominant singularity on the real axis is a pole with multiplicity 1, a case which we will frequently have/assume, we get

$$x(n) \approx K_X R^{-n-1}. \tag{1.35}$$

In this dissertation, we will frequently find pgf's which are only implicitly defined. If an explicit expression can be found, Darboux's theorem can be used to obtain the tail probabilities. If no explicit expression can be found however, we cannot easily find an expression (1.33) and thus we cannot straightforwardly use Darboux's theorem. To cope with implicitly defined pgf's, we will make use of the following theorem:

Theorem 1.3 (Bender's theorem) *Assume that the series $x(z) = \sum_{n=0}^{\infty} x(n) z^n$ with nonnegative coefficients satisfies $F(z, x(z)) = 0$. Suppose there exists real numbers $r > 0$ and $s > x(0)$ such that*

1. *for some $\delta > 0$, $F(z, w)$ is analytic whenever $|z| < r + \delta$ and $|w| < s + \delta$;*
2. *$F(r, s) = F_w(r, s) = 0$;*
3. *$F_z(r, s) \neq 0$, and $F_{ww}(r, s) \neq 0$; and*
4. *if $|z| \leq r$, $|w| \leq s$, and $F(z, w) = F_w(z, w) = 0$, then $z = r$ and $w = s$.*

Then

$$x(n) \sim \left(\frac{r F_z}{2\pi F_{ww}} \right)^{1/2} n^{-3/2} r^{-n}, \tag{1.36}$$

for $n \to \infty$ and where the partial derivatives F_z and F_{ww} are evaluated at $z = r, w = s$.

Remark

The techniques used in this dissertation to calculate the tail probabilities will not be (mathematically) correct for all possible pgf's. Therefore we will sometimes restrict the pgf's for which the obtained tail probabilities are correct. However, we do not think that this restricts the possible application of our results much, since our results are correct for most 'typical' forms of generating functions (such as exponential or rational functions). So one should be careful with using the techniques in this dissertation (and in queueing theory in general) for 'special' functions (and respective stochastic variables).

1.9 Overview of this dissertation

We will end this first chapter by briefly describing the contents of the following chapters. In chapter 2, we analyze a two-class priority queue where the service times of all units are equal to 1 slot. Therefore, no subdivision of the priority scheduling discipline (as described in subsection 1.7.4) has to be made. In chapters 3, 4 and 5 we will analyze two-class priority queues with (more) general service times and a non-preemptive, preemptive resume and preemptive repeat priority scheduling respectively. Finally, we will draw some conclusions in chapter 6.

Chapter 2

Single-slot service times

In this chapter, we analyze a priority queue as described in section 1.7 with two priority classes and deterministic service times of one slot.

Deterministic service times equal to the slot length is a typical type of service time frequently used in *discrete-time* queueing models. This is due to two main reasons. Firstly, it is the easiest model to analyze. This is mainly due to the fact that a unit will leave the system at the end of the slot that it has commenced service. Secondly, constant service times frequently occur in practice. For example in telecommunications, this model is extremely useful for analyzing the performance of switches in an ATM (Asynchronous Transfer Mode) context ([De Prycker 1991]). The main feature of the ATM technology is the fact that all cells - in an ATM context, the information units are called *cells* - have the same length (53 bytes) and the time necessary to transmit a cell is taken as the slot length. We will follow the ATM terminology in the remainder of this chapter.

Discrete-time priority queues with one slot service time and no correlation between the arrival processes of the different priority classes have been analyzed in [Hashida and Takahashi 1991, Takine et al. 1994b, Choi et al. 1998a, Shakkottai and Srikant 2001, Xabier Albizuri et al. 2003] and [Mehmet Ali and Song 2004]. Hashida and Takahashi [1991] analyze a two-class priority system, where the high-priority arrivals and low-priority arrivals are governed by a two-state Markov modulated Batch Bernoulli Process and a Batch Bernoulli Process (or vice-versa). Conservation laws and mean delay are found. In [Takine et al. 1994b], a two-class priority system is investigated. The numbers of per-slot arriving high-priority cells are governed by an underlying Markov chain and the numbers of per-slot low-priority arrivals are i.i.d. Using matrix analytic techniques, moments of high-priority, low-priority and total system contents and moments of high-priority and low-priority delay are calculated. In [Choi et al. 1998a], a two-class priority queue with train arrivals is analyzed. Both priority arrival streams are modeled as fixed-length trains, i.e.,

when the first cell of a train of length m arrives in a slot, $m-1$ cells will arrive in the next $m-1$ slots at a rate of one cell per slot. Using pgf's, it is shown how moments of the queue length and waiting time of cells are calculated. In [Shakkottai and Srikant 2001] bounds for the delay distribution are given in a multi-server queue with a rather general arrival process. Xabier Albizuri et al. [2003] study the delay of the low-priority traffic in a multi-server queue by assuming that the number of servers available for the low-priority traffic is variable (depending on the number of high-priority cells served at the time). Finally, Mehmet Ali and Song [2004] analyze a queue with the arrival process existing out of a number of two-state Markovian sources and by using pgf's.

In [Sidi and Segall 1983, Chang and Harn 1992, Khamisy and Sidi 1992, Laevens and Bruneel 1998, Walraevens and Bruneel 1999] and [Walraevens et al. 2003c], discrete-time priority queues with one slot service time and (some sort of) correlation between the arrival processes of the different priority classes are investigated. In [Sidi and Segall 1983] and [Khamisy and Sidi 1992], priority queueing systems with a general number of priority classes are analyzed. In [Sidi and Segall 1983], the number of arrivals is i.i.d. from slot-to-slot. Departing cells from one priority queue can leave the system or can be switched to another priority queue. In [Khamisy and Sidi 1992], the distribution of the number of per-slot arrivals depends on the state of a two-state Markov chain. In both papers [Sidi and Segall 1983, Khamisy and Sidi 1992], the joint pgf of the system contents of all classes is calculated. From this pgf, the mean system contents and - through (the discretized version of) Little's law (see [Little 1961, Fiems and Bruneel 2002]) - the mean delay of all classes are derived. In [Chang and Harn 1992, Laevens and Bruneel 1998], a two-class multiserver queue is analyzed with the number of arrivals i.i.d. from slot-to-slot. The joint pgf of the system contents of both classes is calculated in both papers (although the analysis in [Chang and Harn 1992] is more tedious as in [Laevens and Bruneel 1998]). The pgf's of the delays of both types of cells are also calculated in [Laevens and Bruneel 1998], while this is not done in [Chang and Harn 1992]. From these pgf's, moments of the analyzed stochastic variables are calculated in both papers. In [Chang and Harn 1992], pmf's are furthermore numerically determined using the Fast Fourier Transform, while these probabilities are analytically approximated for high values of the stochastic variable (tail probabilities) in [Laevens and Bruneel 1998]. Walraevens and Bruneel [1999] and Walraevens et al. [2003c] basically discussed the single-server variant of [Laevens and Bruneel 1998].

In this chapter, we discuss the analysis of a two-class ATM priority queue, as in [Walraevens and Bruneel 1999] and in [Walraevens et al. 2003c]. Although the model we study is a special case of the model in [Laevens and Bruneel 1998] and the models discussed further in this dissertation, we feel it is useful to show the analysis of this simplified queueing model first in full detail. This will permit us to show the techniques used throughout this dissertation in a relatively simple setting.

In section 2.1, we will analyze the system contents. Furthermore, we will

briefly describe corresponding results for the queue contents and unfinished work in sections 2.2 and 2.3. In section 2.4, we will concentrate on the cell delays of both classes and briefly describe some results for the waiting times in section 2.5. Some numerical examples will finally be shown in section 2.6, before giving some concluding remarks in section 2.7.

2.1 System contents

2.1.1 Calculation of the joint pgf $U(z_1, z_2)$

In this subsection, we concentrate on the effect of the priority scheduling discipline on the joint probability generating function of the steady-state system contents of both classes. As discussed in subsection 1.7.2, the number of cell arrivals of class-j in slot k is denoted by $a_{j,k}$, $j = 1, 2$. The joint pgf of $a_{1,k}$ and $a_{2,k}$ is denoted by $A(z_1, z_2)$.

We denote the system contents of class-j at the beginning of slot k by $u_{j,k}$ $(j = 1, 2)$ and the total system contents at the beginning of slot k by $u_{T,k}$. Furthermore, the joint pgf of $u_{1,k}$ and $u_{2,k}$ is denoted by $U_k(z_1, z_2)$, i.e.,

$$U_k(z_1, z_2) \triangleq \mathrm{E}[z_1^{u_{1,k}} z_2^{u_{2,k}}]. \tag{2.1}$$

The system contents of both types of cells are characterized by the following system equations:

$$u_{1,k+1} = [u_{1,k} - 1]^+ + a_{1,k}; \tag{2.2}$$

$$u_{2,k+1} = \begin{cases} [u_{2,k} - 1]^+ + a_{2,k} & \text{if } u_{1,k} = 0 \\ u_{2,k} + a_{2,k} & \text{if } u_{1,k} > 0 \end{cases}, \tag{2.3}$$

where $[.]^+$ denotes the maximum of the argument and 0. Equation (2.2) follows from the observation that class-1 cells are not influenced by class-2 cells. So, when there are class-1 cells in the system at the beginning of slot k, a class-1 cell is served during slot k. The class-1 system contents at the beginning of slot $k+1$ is thus the superposition of the cells in the queue at the beginning of slot k and the class-1 cells arriving during slot k. A class-2 cell on the other hand can only be served, if there are no class-1 cells in the system, i.e., if $u_{1,k} = 0$. This leads to expression (2.3).

Calculation of the joint pgf of the system contents of both classes at the beginning of slot $k + 1$, yields

$$U_{k+1}(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k+1}} z_2^{u_{2,k+1}}\right] \tag{2.4}$$

$$= \mathrm{E}\left[z_1^{u_{1,k+1}} z_2^{u_{2,k+1}} \{u_{1,k} = 0\}\right] + \mathrm{E}\left[z_1^{u_{1,k+1}} z_2^{u_{2,k+1}} \{u_{1,k} > 0\}\right], \tag{2.5}$$

with $\mathrm{E}[X\{Y\}] \triangleq \mathrm{E}[X|Y]\mathrm{Prob}[Y]$. Using the system equations (2.2) and (2.3), we form the following relation between $U_{k+1}(.,.)$ and $U_k(.,.)$

$$U_{k+1}(z_1,z_2) = \frac{A(z_1,z_2)\left\{\begin{array}{l} z_2U_k(z_1,z_2) + (z_1-z_2)U_k(0,z_2) \\ +z_1(z_2-1)U_k(0,0) \end{array}\right\}}{z_1z_2}. \tag{2.6}$$

Notice that

$$U_k(0,z_2) = \mathrm{E}\left[z_2^{u_{2,k}}\{u_{1,k}=0\}\right], \tag{2.7}$$

and

$$U_k(0,0) = \mathrm{Prob}\left[u_{1,k}=u_{2,k}=0\right], \tag{2.8}$$

by definition.

Since we are interested in the steady-state distribution of the system contents, we define $U(z_1,z_2)$ as

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

Applying this limit in equation (2.6), we find the following expression for $U(z_1,z_2)$,

$$U(z_1,z_2) = A(z_1,z_2)\frac{(z_1-z_2)U(0,z_2)+z_1(z_2-1)U(0,0)}{z_2(z_1-A(z_1,z_2))}. \tag{2.9}$$

There are two quantities yet to be determined in the right-hand side of equation (2.9), namely the function $U(0,z_2)$ and the constant $U(0,0)$. Applying Rouché's theorem, it can be proven (see section A.1 and A.2 in the appendix for this proof) that for a given value of z_2 ($|z_2|<1$), the equation $z_1 = A(z_1,z_2)$ has one solution in the unit circle for z_1, which will be denoted by $Y(z_2)$ in the remainder, and which is implicitly defined by $Y(z) \triangleq A(Y(z),z)$. Since $Y(z_2)$ is a zero of the denominator of the right-hand side of (2.9) and since $U(z_1,z_2)$ is analytic for $|z_1|<1$ and $|z_2|<1$ - since $U(z_1,z_2)$ is a pgf - $Y(z_2)$ must also be a zero of the numerator. We thus find

$$U(0,z_2) = U(0,0)\frac{Y(z_2)(z_2-1)}{z_2-Y(z_2)}. \tag{2.10}$$

Substituting this expression in equation (2.9) yields

$$U(z_1,z_2) = U(0,0)\frac{A(z_1,z_2)(z_1-Y(z_2))(z_2-1)}{(z_1-A(z_1,z_2))(z_2-Y(z_2))}. \tag{2.11}$$

Finally, $U(0,0)$ can be found by applying the normalization condition $U(1,1) = 1$. Substituting z_1 and z_2 by 1 in (2.11) and using de l' Hôpital's rule gives the expected result for the probability of having an empty system:

$$U(0,0) = 1 - \lambda_T, \tag{2.12}$$

with $\lambda_T = \left.\frac{dA(z,z)}{dz}\right|_{z=1} = A_T'(1)$ the total arrival rate (as defined in subsection 1.7.2). Using (2.12) in expression (2.11), we finally get the following expression:

$$U(z_1,z_2) = (1-\lambda_T)\frac{A(z_1,z_2)(z_1 - Y(z_2))(z_2-1)}{(z_1 - A(z_1,z_2))(z_2 - Y(z_2))}. \tag{2.13}$$

2.1.2 The function $Y(z)$

It can be proved that $Y(z)$ is a pgf and as a result is analytic inside the unit disk. It is however not easy to show - at this time - which stochastic variable $Y(z)$ is the pgf of, and we will therefore postpone this explanation until after the analysis of the cell delay.

Notice furthermore that the function $Y(z) = A(Y(z),z)$ is only implicitly defined, and that it can only be explicitly calculated for specific arrival processes. In the case of the binomial arrival process discussed in subsection 1.7.2 (see formula (1.8)) with $N = 2$ for instance, $Y(z)$ can be explicitly calculated as

$$Y(z) = \frac{2 - 2\lambda_1 + \lambda_1^2 + \lambda_1\lambda_2(1-z) - 2\sqrt{1 - 2\lambda_1 + \lambda_1^2 + \lambda_1\lambda_2(1-z)}}{\lambda_1^2}. \tag{2.14}$$

For the binomial arrival process with $N \to \infty$, we find

$$Y(z) = e^{\lambda_1(Y(z)-1)}e^{\lambda_2(z-1)}, \tag{2.15}$$

which is a transcendental equation with respect to $Y(z)$. An explicit expression for $Y(z)$ cannot easily be obtained in this case. We will however show further that this is not a problem for calculating the performance measures, such as moments and tail probabilities.

2.1.3 The marginal pgf $U_T(z)$

From equation (2.13), we easily obtain an expression for the pgf $U_T(z)$ describing the total system contents

$$U_T(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{T,k}}\right] \tag{2.16}$$

$$=U(z,z) \tag{2.17}$$

$$=(1-\lambda_T)\frac{A_T(z)(z-1)}{z-A_T(z)}. \tag{2.18}$$

This is the same pgf as the pgf of the system contents of a *single-class* buffer system with $A_T(z)$ the pgf of the number of per-slot arrivals, single-slot service times and with a FIFO (First In First Out) scheduling discipline (see Bruneel and Kim [1993]). Even more generally, this is the same pgf as the pgf of the *total* system contents of a buffer system with $A_T(z)$ the pgf of the total number of per-slot arrivals, single-slot service times and with a (general) *work-conserving* scheduling discipline. Indeed, when the service times of all cells are equal to one slot, the total system contents is independent of the chosen scheduling discipline, as long as it is work-conserving, i.e., as long as the server serves cells when the system is non-empty.

2.1.4 The marginal pgf $U_1(z)$

We furthermore calculate the pgf $U_1(z)$ of the system contents of class-1 cells as follows

$$U_1(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{1,k}}\right] \tag{2.19}$$

$$=U(z,1). \tag{2.20}$$

By substituting z_2 by 1 in expression (2.13) and using de l'Hôpital's rule, we get

$$U_1(z) =\frac{1-\lambda_T}{1-Y'(1)}\frac{A_1(z)(z-1)}{z-A_1(z)}. \tag{2.21}$$

The first derivative of $Y(z)$ evaluated in 1 can be found as follows. Taking the first derivative of both sides of $Y(z) = A(Y(z), z)$ yields

$$Y'(z) =A^{(1)}(Y(z),z)Y'(z) + A^{(2)}(Y(z),z) \tag{2.22}$$

$$=\frac{A^{(2)}(Y(z),z)}{1-A^{(1)}(Y(z),z)}, \tag{2.23}$$

with $A^{(j)}(x,y) \triangleq \left.\dfrac{\partial A(z_1,z_2)}{\partial z_j}\right|_{z_1=x,z_2=y}$, $j = 1, 2$. Substituting z by 1 - and $Y(1)$ by 1 - in (2.23) gives $Y'(1)$ as a function of the arrival rates:

$$Y'(1) = \frac{\lambda_2}{1-\lambda_1}. \tag{2.24}$$

Substituting expression (2.24) in expression (2.21), we finally find

$$U_1(z) = (1-\lambda_1)\frac{A_1(z)(z-1)}{z-A_1(z)}. \tag{2.25}$$

From this expression, we see that the system contents of class-1 cells is not influenced by class-2 cells and furthermore that its pgf has the same structure as $U_T(z)$. This is of course due to the priority scheduling discipline: the class-1 system contents is not influenced by the amount of arriving class-2 cells.

2.1.5 The marginal pgf $U_2(z)$

Finally, we calculate the pgf of the system contents of class-2 cells, denoted by $U_2(z)$, from expression (2.13) as follows

$$U_2(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{2,k}}\right] \tag{2.26}$$

$$= U(1,z) \tag{2.27}$$

$$= (1-\lambda_T)\frac{A_2(z)(z-1)(Y(z)-1)}{(z-Y(z))(A_2(z)-1)}. \tag{2.28}$$

Special case: uncorrelated numbers of per-slot class-1 and class-2 arrivals

In the special case that the numbers of arrivals of class-1 and class-2 cells are uncorrelated, i.e. $A(z_1,z_2) = A_1(z_1)A_2(z_2)$, we can analyze the system contents of class-2 cells in an alternative way. Since class-2 cells can only be served when there are no class-1 cells in the system, we can model the system, with respect to class-2 cells, in terms of a system with server interruptions. The server is blocked for class-2 cells if there are class-1 cells waiting to be sent, and it is available if there are none. We can then calculate the pgf's of the duration of busy and idle periods of class-1 cells, i.e., the time period during which there are class-1 cells in the system (i.e., $u_1 > 0$, with u_1 the steady-state system contents at the beginning of a random slot) and the time period during which there are no such cells (i.e., $u_1 = 0$), respectively. The duration of the idle period is geometrically distributed with parameter $A_1(0)$, since the probability that the idle period lasts an additional slot is equal to the probability that no class-1 cells arrive during that slot, i.e., equal to $A_1(0)$. The pgf of the duration of the idle period is thus given by

$$I(z) = \frac{(1-A_1(0))z}{1-A_1(0)z}. \tag{2.29}$$

The analysis of the busy period is a bit more involved, and can be found in Bruneel and Kim [1993] for a general service time distribution. In case of deterministic service times of one slot, its pgf $B(z)$ is implicitly given by the following formula:

$$B(z) = \frac{A_1(z((1 - A_1(0))B(z) + A_1(0))) - A_1(0)}{1 - A_1(0)}. \tag{2.30}$$

Note that the lengths of consecutive busy and idle periods are statistically independent. It is clear that when the system is busy with respect to class-1 cells, it is blocked for class-2 cells. Therefore, with respect to class-2 cells, the system can be modeled as a single-server buffer with server interruptions, for which the lengths of consecutive available and blocked periods are i.i.d. and their respective pgf's are given by equation (2.29) and (2.30) respectively. Such a queueing system has already been analyzed in Bruneel [1983]. Translating the results from this analysis to our case, the pgf of the system contents of class-2 cells becomes

$$U_2(z) = (1 - \lambda_T)\frac{A_2(z)(z-1)}{1 - A_2(z)}\frac{1 - A_2(z)\left[A_1(0) + (1 - A_1(0))B(A_2(z))\right]}{z - A_2(z)\left[A_1(0) + (1 - A_1(0))B(A_2(z))\right]}. \tag{2.31}$$

Defining

$$X(z) \triangleq A_2(z)\left[A_1(0) + (1 - A_1(0))B(A_2(z))\right], \tag{2.32}$$

this leads to

$$U_2(z) = (1 - \lambda_T)\frac{A_2(z)(z-1)}{z - X(z)}\frac{1 - X(z)}{1 - A_2(z)}. \tag{2.33}$$

Combining (2.32) and (2.30), $X(z)$ is also implicitly given by

$$X(z) = A_1(X(z))A_2(z). \tag{2.34}$$

Equations (2.28) and (2.33) lead to the same result for $U_2(z)$, when $X(z) = Y(z)$. This is indeed the case when the numbers of class-1 and class-2 arrivals during a slot are uncorrelated.

2.1.6 Calculation of moments

The moments of the total, class-1 and class-2 system contents are calculated by taking the necessary derivatives of the respective pgf's and evaluating for $z = 1$ (as explained in subsection 1.8.2). We will explicitly show the expressions of the *means*. Higher (central) moments can also be calculated straight-forwardly,

but expressions are not shown here, although we will show some figures of variances in section 2.6.

The mean total system contents is given by (using expression (2.18))

$$\mathrm{E}[u_T] = \frac{\lambda_T}{2} + \frac{\mathrm{Var}[a_T]}{2(1-\lambda_T)}. \tag{2.35}$$

The mean class-1 system contents is given by (using expression (2.25))

$$\mathrm{E}[u_1] = \frac{\lambda_1}{2} + \frac{\mathrm{Var}[a_1]}{2(1-\lambda_1)}. \tag{2.36}$$

The calculation of the mean class-2 system contents is a bit more involved, because of the appearance of $Y(z)$ in the expression (2.28) of $U_2(z)$. As mentioned in subsection 2.1.2, the function $Y(z)$ can only be explicitly found in case of some simple arrival processes. Its derivatives for $z = 1$, necessary to calculate the moments of the system contents and the cell delay, on the contrary, can be calculated in closed-form. This is because we know $Y(1) = 1$, since $Y(z)$ is a pgf. For example, $Y'(1)$ is given by expression (2.24). Finally, taking the first derivative of (2.28), substituting z by 1 and using expression (2.24) yields

$$\mathrm{E}[u_2] = \frac{\lambda_2}{2} + \frac{\mathrm{Var}[a_2] + 2\mathrm{Cov}[a_1, a_2]}{2(1-\lambda_T)} + \frac{\lambda_2 \mathrm{Var}[a_1]}{2(1-\lambda_T)(1-\lambda_1)}, \tag{2.37}$$

for the mean system contents of class-2 cells, with $\mathrm{Cov}[X, Y]$, the covariance between variables X and Y (as defined in subsection 1.8.2).

Since $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$, and thus

$$\mathrm{Var}[a_T] = \mathrm{Var}[a_1] + \mathrm{Var}[a_2] + 2\mathrm{Cov}[a_1, a_2], \tag{2.38}$$

it is easily verified that equations (2.35), (2.36) and (2.37) satisfy $\mathrm{E}[u_T] = \mathrm{E}[u_1] + \mathrm{E}[u_2]$ (which is expected since $u_T = u_1 + u_2$).

2.1.7 Calculation of tail probabilities

From the pgf's of the total, class-1 and class-2 system contents, (approximations of) the tail probabilities can be derived using Darboux's (Theorem 1.1) or Bender's theorem (Theorem 1.3).

Note that we assume for the reasoning in the remainder that the pgf's of the arrival processes ($A_T(z)$, $A_1(z)$ and $A_2(z)$) and their derivatives go to infinity for z equal to their radii of convergence or for $z \to \infty$ (which is correct for most 'normally' applied arrival distributions). For most pgf's that do not

fulfil this assumptions, the reasoning in this subsection can be adjusted, but this is not the main topic of this dissertation. Our goal in this subsection is twofold: firstly, we want to show that tail probabilities can be calculated - in priority queues - from the obtained pgf's. Secondly, we want to show that the tail probabilities are not necessarily geometrically (or exponentially) decaying, even for 'simple' (pgf's of the) arrival processes.

Total system contents

First we concentrate on the total system contents. From (2.18), it can be seen that the singularities of $U_T(z)$ are composed of the zeros of $z - A_T(z)$ and the (possible) singularities of $A_T(z)$. Under the assumptions mentioned in the previous paragraph, we first prove that the dominant singularity of $U_T(z)$ is a pole with multiplicity 1 and is a zero of $z - A_T(z)$.

From Vivanti's theorem (theorem 1.2), we know that the dominant singularity lies on the positive real axis. We first look at the zeros of $f(z) \triangleq z - A_T(z)$. Its smallest zero on the positive real axis is $z = 1$. Since $f'(1) = 1 - \lambda_T > 0$, this is a zero with multiplicity 1. This is however not a pole of $U_T(z)$ since pgf's remain finite in $z = 1$. Starting from $z = 1$, we look for the next zero of $f(z)$ by increasing z. It is seen that $f(z) > 0$ at first (since $f(1) = 0$ and $f'(1) > 0$). However since $A'_T(z)$ is a strictly increasing function, $f'(z) = 1 - A'_T(z)$ is a strictly decreasing function. Therefore $f(z)$ reaches a maximum for $z = z_*$ (with $f'(z_*) = 0$) and then decreases again. For a certain z_T, $f(z)$ equals zero (again) and $f'(z_T) < 0$. Therefore z_T is a zero with multiplicity 1 of $z - A_T(z)$. Since z_T is inside the region of convergence of $A_T(z)$, z_T is smaller than the (possible) dominant singularity of $A_T(z)$ and is thus the dominant singularity of $U_T(z)$.

For example, if $A_T(z) = (1 - \lambda_T(1-z)/N)^N$, z and $A_T(z)$ are shown in Figure 2.1 for real positive values of z (for $N = 16$ and $\lambda_T = 0.5$). $z - A_T(z)$ has two zeros on the real positive axis. The first one is $z = 1$ and the second one is z_T. $A_T(z)$ is smaller than z for $z \in]1, z_T[$ leading to a positive $f(z) = z - A_T(z)$ for z in that particular range (as described above).

So, in the neighborhood of its dominant pole z_T, we can approximate $U_T(z)$ by

$$U_T(z) \approx \frac{K_T}{z_T - z}, \tag{2.39}$$

since z_T is a single pole of $U_T(z)$. K_T can be found by substituting $z = z_T$ in (2.39) and using the expression (2.18) for $U_T(z)$:

$$K_T = \lim_{z \to z_T} U_T(z)(z_T - z) \tag{2.40}$$

$$= (1 - \lambda_T) z_T (z_T - 1) \lim_{z \to z_T} \frac{z_T - z}{z - A_T(z)} \tag{2.41}$$

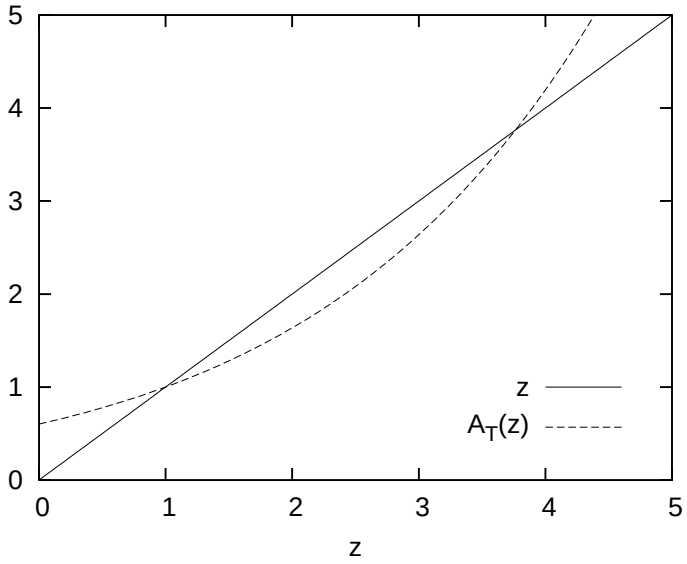


Figure 2.1: The functions z and $A_T(z)$ for z real and positive

$$= \frac{(1-\lambda_T)z_T(z_T-1)}{A'_T(z_T)-1}, \tag{2.42}$$

where we have used de l'Hôpital's rule in the last step. Using Darboux's theorem (or more precisely equations (1.33)-(1.34) with $\omega = 1$) the coefficients of the power series $U_T(z)$ can be found from expression (2.39):

$$u_T(n) \triangleq \text{Prob}[u_T = n] \tag{2.43}$$

$$\approx K_T z_T^{-n-1}, \tag{2.44}$$

for large enough n. Substituting (2.42) in this expression yields

$$u_T(n) \approx \frac{(1-\lambda_T)\,(z_T-1)\,z_T^{-n}}{A'_T(z_T)-1}. \tag{2.45}$$

A quantity of practical interest is the probability that the total system contents exceeds a bound L. We find

$$\text{Prob}[u_T > L] \approx \frac{u_T(L)}{z_T-1} \tag{2.46}$$

$$= \frac{(1-\lambda_T)z_T^{-L}}{A'_T(z_T)-1}. \tag{2.47}$$

Expression (2.46) is found by noting that

$$\sum_{n=0}^{\infty} \text{Prob}[u_T > n] z^n = \frac{U_T(z) - 1}{z - 1}. \tag{2.48}$$

By using Darboux's theorem on this expression - and since this expression inherits the singularities of $U_T(z)$ - expression (2.46) is found.

Note that we could have made the approximation of $U_T(z)$ in (2.39) more accurate by replacing the constant K_T with a function $G(z)$. Since $G(z_T) = K_T$ though and since $G(z_T)$ is ultimately the only value of $G(z)$ that is important for the calculation of the tail probabilities (see Darboux's theorem), we have directly replaced $G(z)$ by K_T in the (approximate) expression of $U_T(z)$. We will do this in all calculations of the tail probabilities in this dissertation.

Class-1 system contents

Since the pgf of the class-1 system contents (expression (2.25)) is similar to the one of the total system contents, the system contents of class-1 cells has an identical tail behavior:

$$u_1(n) \triangleq \text{Prob}[u_1 = n] \tag{2.49}$$

$$\approx \frac{(1 - \lambda_1)(z_H - 1) z_H^{-n}}{A_1'(z_H) - 1}, \tag{2.50}$$

for large enough n, with z_H the dominant singularity on the positive real axis of $U_1(z)$, i.e., z_H is a zero of $z - A_1(z)$.

The probability that the system contents of class-1 exceeds a bound L is given by

$$\text{Prob}[u_1 > L] \approx \frac{(1 - \lambda_1) z_H^{-L}}{A_1'(z_H) - 1}. \tag{2.51}$$

The function $Y(z)$

The tail behavior of the system contents of class-2 cells is a bit more involved, since it is not a priori clear what the dominant singularity is of $U_2(z)$. This is due to the occurrence of the function $Y(z)$ in (2.28), which is only implicitly defined.

First we take a closer look at that function $Y(z)$ on the (positive) real axis (see also section A.3.2 for more details). The first derivative of $Y(z)$ is - as already

mentioned in subsection 2.1.4 - given by

$$Y'(z) = \frac{A^{(2)}(Y(z), z)}{1 - A^{(1)}(Y(z), z)}. \tag{2.52}$$

Consequently, $Y(z)$ has a singularity, denoted as z_B, where the denominator of $Y'(z)$ becomes 0, i.e., $A^{(1)}(Y(z_B), z_B) = 1$. Note that $Y(z_B)$ is finite (for more details see section A.3.2).

As will be proven later on $Y(z)$ is a pgf and thus can be written as a power series with non-negative coefficients:

$$Y(z) = \sum_{n=0}^{\infty} y(n) z^n, \tag{2.53}$$

thus with $y(n)$ a pmf. Since $z_1 = Y(z_2)$ satisfies the equation $z_1 - A(z_1, z_2) = 0$, we will make use of Bender's theorem (Theorem 1.3) to obtain an approximation for the $y(n)$ for n sufficiently high.

To be able to use Bender's theorem, we have to check the four assumptions in this theorem. This is not always an easy task (in general). Therefore, we will not check them for general $A(z_1, z_2)$, but we will - as an example - show that the four assumptions are valid for $A(z_1, z_2)$ equal to the two-dimensional binomial pgf defined in the previous chapter, i.e., for

$$A(z_1, z_2) = \left(1 + \sum_{j=1}^{2} \frac{\lambda_j}{N}(z_j - 1)\right)^N. \tag{2.54}$$

1. for some $\delta > 0$, $z_1 - A(z_1, z_2)$ is analytic whenever $|z_2| < z_B + \delta$ and $|z_1| < Y(z_B) + \delta$: indeed, z_1 and $A(z_1, z_2)$ are analytic in the complete complex plane for our example.

2. $Y(z_B) - A(Y(z_B), z_B) = 1 - A^{(1)}(Y(z_B), z_B) = 0$: this follows directly from the definition of $Y(z)$ and z_B.

3. $A^{(2)}(Y(z_B), z_B) \neq 0$, and $A^{(11)}(Y(z_B), z_B) \neq 0$. For our example, it is easily seen that this holds true for $N > 1$, but for $N = 1$ it is seen that $A^{(11)}(Y(z), z) = 0$ for all z. Note that for $N = 1$ only one packet can enter the queue during a slot - which is served at the beginning of the next slot - and thus there is no queueing at all. The tail probabilities of the system contents are thus zero.

4. if $|z_2| \leq z_B$, $|z_1| \leq Y(z_B)$, and $z_1 - A(z_1, z_2) = 1 - A^{(1)}(z_1, z_2) = 0$, then $z_2 = z_B$ and $z_1 = Y(z_B)$. $z_1 - A(z_1, z_2) = 1 - A^{(1)}(z_1, z_2) = 0$ is in our

example equal to

$$\begin{cases} z_1 - \left(1 + \sum_{j=1}^{2} \frac{\lambda_j}{N}(z_j - 1)\right)^N = 0 \\ 1 - \lambda_1 \left(1 + \sum_{j=1}^{2} \frac{\lambda_j}{N}(z_j - 1)\right)^{N-1} = 0 \end{cases} . \tag{2.55}$$

This set of equations has $N-1$ solutions $(z_1^{(m)}, z_2^{(m)})$:

$$\begin{cases} z_1^{(m)} = \left(\frac{e^{j2\pi m}}{\lambda_1^N}\right)^{1/(N-1)} \\ z_2^{(m)} = \frac{N}{\lambda_2}\left[\frac{N-1}{N}\left(\frac{e^{j2\pi m}}{\lambda_1}\right)^{1/(N-1)} - 1 + \frac{\lambda_T}{N}\right] \end{cases} , \tag{2.56}$$

with $m = 0, \ldots, N-2$. Note that the solution $z_1^{(0)}, z_2^{(0)}$ equals $(Y(z_B), z_B)$. Since $-1 + \lambda_T/N < 0$ it is seen that $|z_2^0| < |z_2^m|$ for all $m > 0$, which means that this assumption is indeed valid.

Note that this last assumption is in general the hardest one to check. Basically, it is equivalent with "$Y(z_2)$ has no other singularities inside and on the circle with radius z_B". Since z_B is one of the dominant singularities, this condition will be true for $|z_2| < z_B$. However, for special arrival processes, $Y(z)$ could have other singularities *on* that circle. More details on conditions for this assumption to be true can be found in Meir and Moon [1989]. They prove for instance that knowing in advance that $y(m)y(n) > 0$ for some $m > n$ with the greatest common devision of m and n equal to 1 is a sufficient condition. However, once again, we stress that it is not always easy to check the assumptions and thus we advice - in respect with Bender's theorem - to "handle with care".

Assuming these 4 conditions are met, Bender's theorem gives the following approximate values of $y(n)$ (for large enough n):

$$y(n) \approx \sqrt{\frac{z_B A^{(2)}(Y(z_B), z_B)}{2\pi A^{(11)}(Y(z_B), z_B)}} n^{-3/2} (z_B)^{-n} . \tag{2.57}$$

An alternative method to obtain $y(n)$ for large enough n is the following: if we find an explicit function for $Y(z_2)$ which is correct in the neighborhood of z_B, we can use Darboux's theorem to obtain the tail probabilities $y(n)$. In Drmota [1997] it is shown (in the more general context of a set of functional equations) that in the neighborhood of z_B, $Y(z)$ is approximately given by

$$Y(z) \approx Y(z_B) - K_Y (z_B - z)^{1/2} . \tag{2.58}$$

This can also be seen from the fact that z_B is a (square-root) branch point of $Y(z)$ (see A.3.2). K_Y can be found from expression (2.58) as follows:

$$K_Y^2 = \lim_{z \to z_B} \frac{(Y(z_B) - Y(z))^2}{z_B - z} \tag{2.59}$$

$$= \lim_{z \to z_B} \left[2 \left(Y(z_B) - Y(z) \right) Y'(z) \right], \tag{2.60}$$

where we have used de 'l Hôpital's rule. Using expression (2.23) for $Y'(z)$, we obtain

$$K_Y^2 = 2A^{(2)}(Y(z_B), z_B) \lim_{z \to z_B} \frac{Y(z_B) - Y(z)}{1 - A^{(1)}(Y(z), z)} \tag{2.61}$$

$$= 2A^{(2)}(Y(z_B), z_B) \lim_{z \to z_B} \frac{Y'(z)}{A^{(11)}(Y(z), z) Y'(z) + A^{(12)}(Y(z), z)}, \tag{2.62}$$

after using de 'l Hôpital's rule once more. Since $Y'(z) \to \infty$ for $z \to z_B$, the last term of the denominator of (2.62) is negligible. This ultimately leads too

$$K_Y = \sqrt{\frac{2A^{(2)}(Y(z_B), z_B)}{A^{(11)}(Y(z_B), z_B)}}. \tag{2.63}$$

Using Darboux's theorem on expression (2.58), we get

$$y(n) = -\frac{K_Y \sqrt{z_B}}{\Gamma(-1/2)} n^{-3/2} z_B^{-n}. \tag{2.64}$$

Using expression (2.63) and the knowledge that $\Gamma(-1/2) = -2\sqrt{\pi}$, we indeed find expression (2.57). Both methods thus lead to the same result.

Class-2 system contents

Since $Y(z)$ appears in the expression (2.28) of $U_2(z)$, z_B is also a singularity of $U_2(z)$. Indeed, taking the first derivative of expression (2.28) yields

$$U_2'(z) = \frac{(1 - \lambda_T) \left\{ \begin{array}{l} (z-1)(1 - Y(z))(z - Y(z)) A_2'(z) \\ + A_2(z)(1 - Y(z))^2 (1 - A_2(z)) \\ - A_2(z)(z-1)^2 (1 - A_2(z)) Y'(z) \end{array} \right\}}{(z - Y(z))^2 (1 - A_2(z))^2}, \tag{2.65}$$

and it is easily seen that this expression goes to infinity as $Y'(z) \to \infty$, or, as $z \to z_B$.

A second potential singularity z_L of $U_2(z)$ on the real axis is given by the positive zero of the denominator $z - Y(z)$, and it is easily proved to be equal

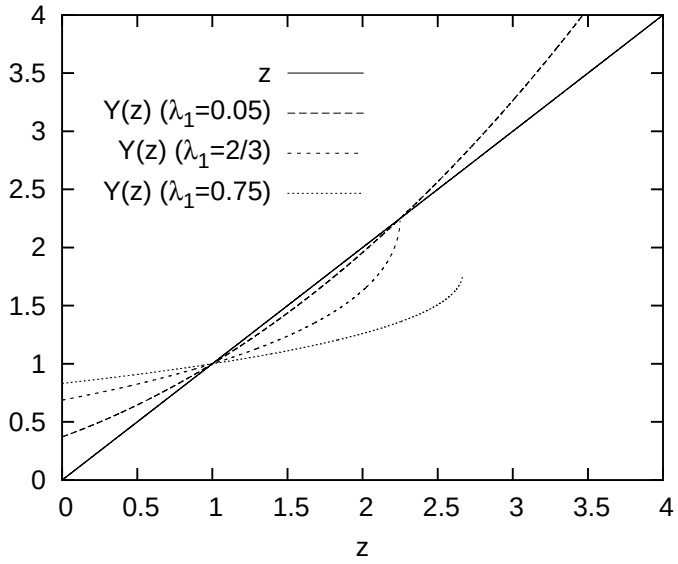


Figure 2.2: Types of behavior of $Y(z)$

to z_T, if z_L exists. Figure 2.2 gives three typical types of behavior of $Y(z)$ for $A(z_1, z_2) = (1 - \lambda_1(1 - z_1)/N - \lambda_2(1 - z_2)/N)^N$, for $N = 2$, $\lambda_T = 0.8$ and $\lambda_1 = 0.05, 2/3$ and 0.75 respectively. For $\lambda_1 = 0.05$, $Y(z)$ intersects z twice (for $z = 1$ and for $z = z_L$), before reaching the branch point (not shown in the Figure). For $\lambda_1 = 2/3$, $Y(z)$ intersects z once in $z = 1$ and equals z in its branch point. When $\lambda_1 = 0.75$ finally, $Y(z)$ intersects z once in $z = 1$ and reaches its branch point before it could intersect z a second time. In this case, no z_L is found, or alternatively, $z - Y(z) \neq 0$ for real $z > 1$ (and for z for which $Y(z)$ exists), see Appendix for more details.

The tail behavior of the system contents of class-2 cells is thus characterized by z_L or z_B, depending on which is the dominant (i.e., smallest) singularity. Furthermore, z_L equals z_T when it is dominant (or equivalently, when it exists). Three situations may thus occur, namely when $z_L = z_T < z_B$, z_L does not exist, and $z_L = z_T = z_B$. We will discuss these three cases separately first, or more precisely we will first study the (approximate) behavior of $U_2(z)$ in the neighborhood of its dominant singularity for the three cases separately.

In the first case, the single pole z_T is dominant and thus

$$U_2(z) \approx \frac{K_2^{(1)}}{z_T - z}, \tag{2.66}$$

for $z \to z_T$. $K_2^{(1)}$ can be calculated by substituting expression (2.28) in the pre-

vious expression and substituting $z = z_T$ (in a similar way as in the calculation of (2.42)). This yields

$$K_2^{(1)} = \frac{(1-\lambda_T)A_2(z_T)\,(z_T-1)^2}{(A_2(z_T)-1)\,(Y'(z_T)-1)}. \tag{2.67}$$

In the second case, i.e., when z_L does not exist, the branch point z_B is dominant. Using the definition of $Y(z)$ ($Y(z) \triangleq A(Y(z),z)$), expression (2.28) can be transformed in a functional equation. Using Bender's theorem, it is then possible to obtain the tail behavior. Alternatively, we first study the behavior of $U_2(z)$ in the neighborhood of z_B. Using expression (2.58) in (2.28), we find

$$U_2(z) \approx (1-\lambda_T)\frac{A_2(z)(z-1)\left(Y(z_B)-K_Y\,(z_B-z)^{1/2}-1\right)}{\left(z-Y(z_B)+K_Y\,(z_B-z)^{1/2}\right)(A_2(z)-1)} \tag{2.68}$$

$$\approx (1-\lambda_T)\frac{\left\{\begin{array}{l} A_2(z)(z-1)\left(Y(z_B)-K_Y\,(z_B-z)^{1/2}-1\right) \\ \times\left(z-Y(z_B)-K_Y\,(z_B-z)^{1/2}\right)\end{array}\right\}}{\left((z-Y(z_B))^2-K_Y^2\,(z_B-z)\right)(A_2(z)-1)}. \tag{2.69}$$

This expression leads to

$$U_2(z) \approx U_2(z_B) - K_2^{(3)}\,(z_B-z)^{1/2}, \tag{2.70}$$

in the neighborhood of z_B - note that we have used the notation $K_2^{(3)}$ instead of (the expected) $K_2^{(2)}$ because we will switch the last two types of tail behavior in the end formula's - with

$$K_2^{(3)} = \frac{(1-\lambda_T)K_Y A_2(z_B)\,(z_B-1)^2}{(A_2(z_B)-1)\,(z_B-Y(z_B))^2}. \tag{2.71}$$

In the third case, z_T and z_B coincide. Again, we will study the behavior of $U_2(z)$ in the neighborhood of this dominant singularity and use Darboux's theorem to calculate the tail probabilities of the class-2 system contents. The approximation of $U_2(z)$ in the neighborhood of z_B is again found by substituting expression (2.58) in expression (2.28):

$$U_2(z) \approx (1-\lambda_T)\frac{A_2(z)(z-1)\left(z_B-K_Y\,(z_B-z)^{1/2}-1\right)}{\left(z-z_B+K_Y\,(z_B-z)^{1/2}\right)(A_2(z)-1)} \tag{2.72}$$

$$\approx(1-\lambda_T)\frac{A_2(z)(z-1)\left(z_B-K_Y\left(z_B-z\right)^{1/2}-1\right)}{\left(z_B-z\right)^{1/2}\left(\left(z_B-z\right)^{1/2}+K_Y\right)\left(A_2(z)-1\right)}, \tag{2.73}$$

where we have also used the fact that $Y(z_B)=z_B$. This leads to the following form of $U_2(z)$ in the neighborhood of its dominant singularity:

$$U_2(z)=\frac{K_2^{(2)}}{\left(z_B-z\right)^{1/2}}, \tag{2.74}$$

with

$$K_2^{(2)}=\frac{(1-\lambda_T)A_2(z_B)\left(z_B-1\right)^2}{K_Y\left(A_2(z_B)-1\right)}. \tag{2.75}$$

Summarizing, $U_2(z)$ can be approximated in the neighborhood of its dominant singularity by:

$$U_2(z)\approx\begin{cases}\dfrac{K_2^{(1)}}{z_T-z} & \text{if } z_L=z_T<z_B\\ \dfrac{K_2^{(2)}}{\left(z_B-z\right)^{1/2}} & \text{if } z_L=z_T=z_B\\ U_2(z_B)-K_2^{(3)}\left(z_B-z\right)^{1/2} & \text{if } z_L \text{ does not exist,}\end{cases} \tag{2.76}$$

where the constants $K_2^{(i)}$ ($i=1,2,3$) are given by expressions (2.67), (2.75) and (2.71) respectively (note that we switched the second and third case). Using Darboux's theorem (see Theorem 1.1), we find the tail probabilities for the three possible cases:

$$u_2(n)\triangleq\text{Prob}[u_2=n] \tag{2.77}$$

$$\approx\begin{cases}\dfrac{(1-\lambda_T)A_2(z_T)\left(z_T-1\right)^2 z_T^{-n-1}}{\left(A_2(z_T)-1\right)\left(Y'(z_T)-1\right)}\\ \dfrac{(1-\lambda_T)A_2(z_B)\left(z_B-1\right)^2 n^{-1/2}z_B^{-n}}{K_Y\sqrt{z_B\pi}\left(A_2(z_B)-1\right)}\\ \dfrac{(1-\lambda_T)K_Y A_2(z_B)\left(z_B-1\right)^2 n^{-3/2}z_B^{-n}}{2\sqrt{\pi/z_B}\left(A_2(z_B)-1\right)\left(z_B-Y(z_B)\right)^2},\end{cases} \tag{2.78}$$

for large enough n, if $z_L=z_T<z_B$, if $z_L=z_T=z_B$ and if z_L does not exist, respectively. The first expression constitutes a typical *geometric* (or *exponential*) tail behavior, while the third expression is a typical *non-geometric* tail behavior. The second expression exhibits a behavior in between the two other cases, and we will thus call this tail behavior of *transition* type.

Finally, the probability that the system contents of class-2 exceeds a bound L is (approximately) given by

$$\text{Prob}[u_2 > L] \approx \frac{u_2(L)}{z_* - 1}, \tag{2.79}$$

with z_* the dominant singularity of $U_2(z)$ and for large enough L. This can be found in the same way as (2.46).

2.2 Queue contents

The *queue contents*, defined as the number of cells in the queue (thus without the one in the server if any), can be easily derived from the system contents. We denote the queue contents of class-j at the beginning of the k-th slot by $q_{j,k}$ ($j = 1, 2$). We then get the following relation between the $q_{j,k}$ and the $u_{j,k+1}$:

$$u_{j,k+1} = q_{j,k} + a_{j,k}, \tag{2.80}$$

for $j = 1, 2$. $u_{j,k+1}$ and $a_{j,k}$ are still defined as the class-j system contents at the beginning of slot $k + 1$ and the number of class-j arrivals during slot k respectively. This relation can be understood as follows: the system contents at the beginning of slot $k + 1$ exists out of the queue contents at the beginning of the previous slot (the possible cell in the server during slot k has left the system at the end of that slot) and the cells that arrived during slot k. Let us denote the joint pgf of the steady-state queue contents by $Q(z_1, z_2)$, i.e.,

$$Q(z_1, z_2) = \lim_{k\to\infty} \text{E}\left[z_1^{q_{1,k}} z_2^{q_{2,k}}\right]. \tag{2.81}$$

By z-transforming the system equations (2.80) and letting $k \to \infty$, we find

$$Q(z_1, z_2) = \frac{U(z_1, z_2)}{A(z_1, z_2)}. \tag{2.82}$$

Substituting $U(z_1, z_2)$ with its expression (2.13) finally yields:

$$Q(z_1, z_2) = (1 - \lambda_T)\frac{(z_1 - Y(z_2))(z_2 - 1)}{(z_1 - A(z_1, z_2))(z_2 - Y(z_2))}. \tag{2.83}$$

From this expression, marginal pgf's, moments and tail probabilities of the total, the class-1 and the class-2 queue contents can be calculated as is done in section 2.1 for the system contents. We will not give the expressions here, because it results in basically the same expressions as in the previous section, but without the factor introduced by the factor $A(z_1, z_2)$ in (2.13).

2.3 Unfinished work

In the case of single-slot service times, the *system contents* and the *unfinished work* are equal. This is because every arriving cell in the queue adds *one* unit to the system contents and adds *one* slot to the unfinished work. So, denoting $W(z_1, z_2)$ as the pgf of the steady-state unfinished work at the beginning of a random slot, or

$$W(z_1, z_2) = \lim_{k \to \infty} \mathrm{E}\left[z_1^{w_{1,k}} z_2^{w_{2,k}}\right], \tag{2.84}$$

with $w_{j,k}$ ($j = 1, 2$) the unfinished work of class-j at the beginning of slot k, $W(z_1, z_2)$ equals (see expression (2.13))

$$W(z_1, z_2) = U(z_1, z_2) \tag{2.85}$$

$$= (1 - \lambda_T) \frac{A(z_1, z_2)(z_1 - Y(z_2))(z_2 - 1)}{(z_1 - A(z_1, z_2))(z_2 - Y(z_2))}. \tag{2.86}$$

From this joint pgf, the same marginal pgf's and performance measures can be calculated as in section 2.1.

Note that the unfinished work and the system contents are equal because of the single-slot service times. Since we will assume more general service times in the other chapters, these two stochastic variables will have to be analyzed separately in the following chapters (although they will obviously still be related).

2.4 Cell delay

In this section, we analyze the *cell delay* of the class-1 and class-2 cells respectively. As stated in chapter 1, the delay of a tagged cell is - in a discrete-time context - defined as the number of slots between the end of its arrival slot and the end of its departure slot. In this section, we derive expressions for the pgf's of the steady-state cell delay of both classes and of the steady-state delay of a random cell, and calculate the related performance measures.

2.4.1 Pgf $D_1(z)$ of the class-1 cell delay

We can analyze the cell delay of class-1 cells - the high-priority cells - as if they were the only type of cells in the system. Indeed, when a class-1 cell arrives it is served before all class-2 cells in the system at that time and thus the high-priority cells do not "see" the low-priority cells. Therefore, the class-1 cell delay is the same as in a corresponding *single-class* system with only class-1 cells arriving. This analysis is (obviously) already done in the past - in a more

general context (see e.g. Bruneel and Kim [1993]), but to set the mind of the reader we will reconstruct this analysis in the setting of this chapter. We tag a class-1 cell and assume that the arrival slot of the tagged cell is the k-th slot. The delay d_1 of this tagged cell is given by

$$d_1 = [u_{1,k} - 1]^+ + f_{1,k}^{(1)} + 1, \tag{2.87}$$

with $u_{1,k}$ the system contents of class-1 at the beginning of slot k and $f_{1,k}^{(1)}$ is defined as the number of class-1 cells that arrive during the arrival slot of the tagged cell, but which have to be served before it (or which arrive during the same slot as the tagged cell but "before" it). Indeed, the tagged cell has to wait in the queue until all class-1 cells that were already in the queue when it arrives (i.e., all cells that were already in the queue at the beginning of its arrival slot and all cells that arrived in front of the tagged cell in its arrival slot) are served. The delay then equals this waiting time augmented with its own service time, which equals 1 in this model. This leads to expression (2.87). Note that we could express this expression also in terms of queue contents instead of system contents (this would lead to a more "simplified" expression in this case because $[u_{1,k} - 1]^+ = q_{1,k}$). This is a matter of choice, and we will throughout this dissertation choose to express the delay as a function of the system contents. Translating expression (2.87) into pgf's yields

$$D_1(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{d_1}\right] \tag{2.88}$$

$$= F_1^{(1)}(z)\left[U_1(z) + (z-1)U_1(0)\right], \tag{2.89}$$

with

$$F_1^{(1)}(z) \triangleq \mathrm{E}\left[z^{f_{1,k}^{(1)}}\right]. \tag{2.90}$$

Notice that we have used the fact that $u_{1,k}$ and $f_{1,k}^{(1)}$ are uncorrelated - because the arrival process is i.i.d. from slot-to-slot - to obtain (2.89). $F_1^{(1)}(z)$ can be calculated by taking into account that an arbitrary cell is more likely to arrive in a larger bulk (e.g. Bruneel and Kim [1993]), yielding - we will describe the calculation of a similar (but more general) pgf in more detail in the following subsection -

$$F_1^{(1)}(z) = \frac{A_1(z) - 1}{\lambda_1(z-1)}. \tag{2.91}$$

Using expressions (2.25) and (2.91) in (2.89), we finally find

$$D_1(z) = \frac{1-\lambda_1}{\lambda_1}\frac{z(A_1(z)-1)}{z - A_1(z)}. \tag{2.92}$$

Notice that $U_1(z)$ (expression (2.25)) and $D_1(z)$ (expression (2.92)) fulfil the following general relation (see Xiong and Bruneel [1993] or Vinck and Bruneel [1995]):

$$U_1(z) = 1 - \lambda_1 + \lambda_1 D_1(z). \tag{2.93}$$

2.4.2 Pgf $D_2(z)$ of the class-2 cell delay

The analysis of the cell delay of class-2, denoted by d_2, is more complicated. Consider a logically equivalent queueing system where all class-1 cells are stored in front of the class-2 cells, and let us tag an arbitrary class-2 cell that arrives in the system. The amount of time it spends in the system equals

$$d_2 = \sum_{j=1}^{[u_{T,k}-1]^+ + f_{T,k}^{(2)}} v_j + 1, \tag{2.94}$$

where slot k is assumed to be the arrival slot of the tagged class-2 cell, $u_{T,k}$ is the total system contents at the beginning of slot k, $f_{T,k}^{(2)}$ is defined as the total number of cells that arrive during the arrival slot of the tagged cell, but which have to be served before it, and v_j is the length of the j-th sub-busy period (from slot k onwards) initiated by the cells already in the queue when the class-2 cell arrives. Note that, since all class-1 and class-2 cells have single-slot service times, it is sufficient to know the total system contents at the beginning of slot k in order to analyze the class-2 cell delay in this model. In the following chapters, the distributions of the class-1 and class-2 service times may be different and as a result $u_{1,k}$ and $u_{2,k}$ will have to be known separately (instead of only their sum $u_{T,k}$).

The notion of a *sub-busy period initiated by a cell* is widely used in queueing analyses. It is - in the context of our analysis - basically defined as follows: the sub-busy period starts at the beginning of the slot the cell enters the server. Assume that at that time instant, m cells are waiting in the queue in front of the tagged class-2 cell (in the logically equivalent queueing system, i.e., all class-1 cells are stored in front of the class-2 cells). The sub-busy period ends at the beginning of the slot where - for the first time - the number of cells waiting before the tagged class-2 cell equals $m-1$, i.e., equals one less than at the beginning of the sub-busy period. In case of a FIFO scheduling, v_j would equal 1 (see e.g. also expression (2.87)). For a priority scheduling, this is not necessarily the case, since new class-1 cells can arrive while the tagged cell is waiting in the queue and these class-1 cells have to be served *before* the tagged cell. More specifically, assume that the tagged cell is stored in the j-th position in the queue at the beginning of the l-th slot. If no class-1 cells arrive during slot l, v_j equals 1. If $a_{1,l}$ (> 0) class-1 cells arrive during this slot on the other hand, the tagged cell will move back to position $j + a_{1,l} - 1$ in the queue at

the beginning of slot $l+1$, since these class-1 cells have to be served before all class-2 cells, and thus before the tagged one.

Since the arrival process is i.i.d. from slot-to-slot it is obvious that the v_j are all i.i.d. stochastic variables. We denote their common pgf by $V(z)$. Since $u_{T,k}$, $f_{T,k}^{(2)}$ and the v_j are all mutually independent variables, z-transforming expression (2.94) yields

$$D_2(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{d_2}\right] \tag{2.95}$$

$$= zF_T^{(2)}(V(z))\frac{U_T(V(z)) + (V(z)-1)U_T(0)}{V(z)}, \tag{2.96}$$

with $F_T^{(2)}(z)$ the pgf of $f_{T,k}^{(2)}$ and with $U_T(z)$ the pgf of the total system contents at the beginning of a random slot. Furthermore, $f_{T,k}^{(2)}$ is the sum of all the class-1 cells that arrive during the same slot as the tagged one, and of the class-2 cells that have arrived before it during its arrival slot. The pgf of $f_{T,k}^{(2)}$ is calculated first. We define

$$f_T^{(2)}(n) \triangleq \text{Prob}\left[f_{T,k}^{(2)} = n\right], \tag{2.97}$$

and

$$\hat{a}(m,n) \triangleq \text{Prob}\left[\hat{a}_1 = m, \hat{a}_2 = n\right], \tag{2.98}$$

with $\hat{a}_j$ the number of class-j arrivals in the arrival slot of a tagged class-2 cell. Taking into account that an arbitrary tagged cell is more likely to arrive in a larger bulk $\hat{a}(m,n)$ is given by

$$\hat{a}(m,n) = \frac{na(m,n)}{\lambda_2}, \tag{2.99}$$

with $a(m,n)$ the probability mass function of the number of class-1 and class-2 arrivals in a random slot. Note that $\hat{a}(m,n)$ and $a(m,n)$ are not equal (see Bruneel and Kim [1993] for more details). When $f_{T,k}^{(2)} = n$, the number of class-1 arrivals is at most n and the total number of arrivals during slot k has to be larger than n, leading to

$$f_T^{(2)}(n) = \sum_{m=0}^{n} \sum_{i=n-m+1}^{\infty} \frac{\hat{a}(m,i)}{i}. \tag{2.100}$$

Substituting expression (2.99) in this expression and taking the z-transform

yields

$$F_T^{(2)}(z) = \frac{A_T(z) - A_1(z)}{\lambda_2(z-1)}. \tag{2.101}$$

Using equations (2.18) and (2.101) in expression (2.96) gives

$$D_2(z) = \frac{1-\lambda_T}{\lambda_2} \frac{z(A_T(V(z)) - A_1(V(z)))}{V(z) - A_T(V(z))}. \tag{2.102}$$

It remains for us to determine $V(z)$. During the first slot of a sub-busy period, class-1 packets arrive which all initiate sub-busy periods of their own, part of the initial sub-busy period. The newly introduced sub-busy periods are stochastically indistinguishable from the initial sub-busy period, so they (all) have the same pgf $V(z)$. Thus, since the length of the initial sub-busy period equals one (the first slot) added with the sum of the lengths of the sub-busy periods (initiated by the class-1 arrivals during the first slot), $V(z)$ is *implicitly* given by

$$V(z) = zA_1(V(z)). \tag{2.103}$$

Expression (2.102) is then further transformed in

$$D_2(z) = \frac{1-\lambda_T}{\lambda_2} \frac{zA_T(V(z)) - V(z)}{V(z) - A_T(V(z))}, \tag{2.104}$$

with $V(z)$ implicitly given by (2.103).

2.4.3 Pgf $D(z)$ of the delay of a random cell

In this subsection, we will derive the pgf $D(z)$ of a *random* cell arriving in the system. Tagging a random arriving cell, it is of class-1 with probability λ_1/λ_T and of class-2 with probability λ_2/λ_T. We thus get

$$D(z) = \frac{\lambda_1}{\lambda_T} D_1(z) + \frac{\lambda_2}{\lambda_T} D_2(z). \tag{2.105}$$

Substituting expressions (2.92) and (2.104) in this expression leads to

$$D(z) = \frac{1-\lambda_1}{\lambda_T} \frac{z(A_1(z)-1)}{z - A_1(z)} + \frac{1-\lambda_T}{\lambda_T} \frac{zA_T(V(z)) - V(z)}{V(z) - A_T(V(z))}. \tag{2.106}$$

2.4.4 The function $Y(z)$ revisited

We mentioned in subsection 2.1.2 that $Y(z)$ - defined as $Y(z) \triangleq A(Y(z), z)$ - is a pgf. In this subsection, we will "define" a stochastic variable with pgf $Y(z)$.

The function $Y(z)$ is the pgf of the stochastic variable y, which is defined as the number of class-2 cells that arrive during a sub-busy period (denoted by v) initiated by a random cell (as defined in subsection 2.4.2), i.e.,

$$y = \sum_{i=1}^{v} a_2^{(i)}, \tag{2.107}$$

with $a_2^{(i)}$ defined as the number of class-2 arrivals during the i-th slot of v. We furthermore define the number of class-1 arrivals during the i-th slot of v by $a_1^{(i)}$. The sub-busy period v exists out of the first slot, needed to serve the cell that initiates the sub-busy period, and $a_1^{(1)}$ sub-busy periods (initiated by the class-1 arrivals during the first slot), all with the same distribution as the original sub-busy period. y is thus given by

$$y = a_2^{(1)} + \sum_{m=1}^{a_1^{(1)}} y_m^{(1)}, \tag{2.108}$$

with $y_m^{(1)}$ the number of class-2 cells that arrive during the sub-busy period initiated by the m-th class-1 cell that arrives during the first slot of v. Naturally, all $y_m^{(1)}$ have the same distribution as y (since the lengths of all sub-busy periods are also i.i.d.) and their pgf is thus indeed given by

$$Y(z) = A(Y(z), z), \tag{2.109}$$

as immediately follows by z-transforming (2.108).

Note that $Y(z)$ not necessarily equals $V(A_2(z))$ as one would expect from the definition of the stochastic variable y. Since v depends on the $a_1^{(i)}$ and since - for each i - $a_1^{(i)}$ and $a_2^{(i)}$ are correlated, v also depends on the $a_2^{(i)}$. Thus, since v and $a_2^{(i)}$ are correlated, $Y(z)$ does not equal $V(A_2(z))$ in general. This will be the case however, when the number of class-1 and class-2 arrivals in a slot are mutually independent. Indeed, when $A(z_1, z_2) = A_1(z_1)A_2(z_2)$, $Y(z)$ and $V(A_2(z))$ are solutions of the same functional equation $x = A_1(x)A_2(z)$ (as can be seen from expressions (2.109) and (2.103) respectively) and thus since they are both pgf's $Y(z) = V(A_2(z))$ in this case.

2.4.5 Calculation of moments

Taking the first derivative of equation (2.92) and substituting $z = 1$ yields the mean delay of a class-1 cell:

$$\mathrm{E}[d_1] = \frac{1}{2} + \frac{\mathrm{Var}[a_1]}{2\lambda_1(1-\lambda_1)}. \tag{2.110}$$

Note that this expression is always at least equal to 1, since $\mathrm{Var}[a_1] \geq \lambda_1(1-\lambda_1)$. This can intuitively be seen as follows: in order for a discrete stochastic variable to have a mean value of $\lambda_1 < 1$ and a minimal variance, it has to be 1 with probability λ_1 and 0 with probability $1-\lambda_1$. This distribution has a variance of $\lambda_1(1-\lambda_1)$ and this is thus the least possible variance of the number of per-slot class-1 arrivals. Note that this was intuitively expected since the service time of a cell is always an integral part of the delay of that cell (and thus $\mathrm{E}[d_1] \geq 1$). Thus the mean delay is at least the mean service time of a cell, which equals 1 in this model.

The mean delay of a class-2 cell is found by taking the first derivative of expression (2.104) and substituting z by 1, yielding

$$\mathrm{E}[d_2] = \frac{1}{2} + \frac{\mathrm{Var}[a_2] + 2\mathrm{Cov}[a_1, a_2]}{2\lambda_2(1-\lambda_T)} + \frac{\mathrm{Var}[a_1]}{2(1-\lambda_T)(1-\lambda_1)}, \tag{2.111}$$

for the mean cell delay of a class-2 cell. This expression is also at least one, since for given mean arrival rates of class-1 and class-2 (λ_1 and λ_2 respectively), the minimal value of $\mathrm{Var}[a_j]$ is $\lambda_j(1-\lambda_j)$ ($j = 1, 2$) and the minimal value of $\mathrm{Cov}[a_1, a_2]$ is $-\lambda_1\lambda_2$. These least possible values occur when one class-1 cell and no class-2 cells arrive during a slot with probability λ_1, no class-1 cells and one class-2 cell arrive with probability λ_2 and no arrivals occur during a slot with probability $1-\lambda_T$.

Finally the mean delay of a random cell equals

$$\mathrm{E}[d] = \frac{\lambda_1}{\lambda_T}\mathrm{E}[d_1] + \frac{\lambda_2}{\lambda_T}\mathrm{E}[d_2] \tag{2.112}$$

$$= \frac{1}{2} + \frac{\mathrm{Var}[a_1] + \mathrm{Var}[a_2] + 2\mathrm{Cov}[a_1, a_2]}{2\lambda_T(1-\lambda_T)}. \tag{2.113}$$

We have used expressions (2.110) and (2.111) to obtain this formula. Note that this expression can (obviously) also be found by taking the first derivative of expression (2.106) and substituting z by 1. Finally, since

$$\mathrm{Var}[a_T] = \mathrm{Var}[a_1 + a_2] \tag{2.114}$$

$$= \mathrm{Var}[a_1] + \mathrm{Var}[a_2] + 2\mathrm{Cov}[a_1, a_2], \tag{2.115}$$

we find

$$\mathrm{E}[d] = \frac{1}{2} + \frac{\mathrm{Var}[a_T]}{2\lambda_T(1-\lambda_T)}. \tag{2.116}$$

This expression is the same expression as in a single-class queue with a FIFO scheduling discipline, or, in other words the mean delay of a *random* cell is independent whether the scheduling discipline is FIFO or there is some type of priority scheduling discipline. This is even extendable to every *work-conserving* scheduling discipline. This can easily be proven by (the discretisized version of) Little's law (see Little [1961] and Fiems and Bruneel [2002]). For each work-conserving scheduling discipline, the distribution of the (total) system contents is identical and thus leads to the same pgf $U_T(z)$ (expression (2.18)), and to a same mean total system contents $\mathrm{E}[u_T]$ (expression (2.35)). Little's law, in our model given by

$$\mathrm{E}[u_T] = \lambda_T \mathrm{E}[d], \tag{2.117}$$

then gives a same mean delay for all scheduling disciplines. Indeed, expressions (2.35) and (2.116) satisfy Little's law.

Little's law is not only valid for the "total" buffer, but can also be used on "parts of the buffer/cells". In a priority context, for instance, Little's law is also valid for each priority class separately. Or, more precisely,

$$\mathrm{E}[u_j] = \lambda_j \mathrm{E}[d_j], \tag{2.118}$$

for $j = 1, 2$. It is easily verified that expressions (2.36)-(2.110) and expressions (2.37)-(2.111) respectively satisfy Little's law.

2.4.6 Calculation of tail probabilities

From the pgf's of the delay of a class-1 cell, a class-2 cell and a random cell, the tail probabilities can be derived using Darboux's theorem (Theorem 1.1) or Bender's theorem (Theorem 1.3), in a similar way as for the system contents (see subsection 2.1.7). Details about the methods used can be found in that subsection. We will here only give a brief overview of the method and obtained results.

Class-1 delay

The dominant singularity of $D_1(z)$ (expression (2.92)) is the same as the one of $U_1(z)$ (also with multiplicity 1), i.e., z_H, the dominant zero of $z - A_1(z)$. In

the neighborhood of this singularity, $D_1(z)$ is given by

$$D_1(z) \approx \frac{(1-\lambda_1)z_H\,(z_H-1)}{\lambda_1\,(A_1'(z_H)-1)\,(z_H-z)}. \tag{2.119}$$

Using Darboux's theorem, we approximate the tail behavior of the delay of class-1 cells by

$$d_1(n) \triangleq \text{Prob}[d_1 = n] \tag{2.120}$$

$$\approx \frac{(1-\lambda_1)\,(z_H-1)\,z_H^{-n}}{\lambda_1\,(A_1'(z_H)-1)}, \tag{2.121}$$

for large enough n. Summing this equation for all $n = D+1, \ldots, \infty$ gives the probability that the delay of a class-1 cell exceeds a bound D:

$$\text{Prob}[d_1 > D] \approx \frac{(1-\lambda_1)z_H^{-D}}{\lambda_1\,(A_1'(z_H)-1)}, \tag{2.122}$$

Sub-busy periods

The tail behavior of the delay of class-2 cells is again a bit more involved because of the appearance of the function $V(z)$ in (2.104), which is only implicitly known by expression (2.103). The first derivative of $V(z)$ is given by

$$V'(z) = \frac{A_1(V(z))}{1 - zA_1'(V(z))}, \tag{2.123}$$

which, similar as for $Y(z)$, indicates that $V(z)$ has a branch point $\hat{z}_B$, with $\hat{z}_B A_1'(V(\hat{z}_B)) = 1$. In the neighborhood of $\hat{z}_B$, $V(z)$ is approximately given by

$$V(z) \approx V(\hat{z}_B) - K_V\,(\hat{z}_B - z)^{1/2}\,, \tag{2.124}$$

with

$$K_V = \sqrt{\frac{2A_1(V(\hat{z}_B))}{\hat{z}_B A_1''(V(\hat{z}_B))}}. \tag{2.125}$$

Using Bender's theorem on the functional equation (2.103) or Darboux's theorem on (2.124) leads to the tail probabilities $v(n)$ of a sub-busy period initiated by a class-1 cell

$$v(n) = \sqrt{\frac{A_1(V(\hat{z}_B))}{2\pi A_1''(V(\hat{z}_B))}}\, n^{-3/2} \hat{z}_B^{-n} \tag{2.126}$$

Class-2 delay

Since $V(z)$ appears in the expression of $D_2(z)$, $\hat{z}_B$ is also a branch-point type singularity of $D_2(z)$. A second singularity of $D_2(z)$ is given by the dominant zero $\hat{z}_L$ of $V(z) - A_T(V(z))$ on the real axis. It is proved to equal $\frac{z_T}{A_1(z_T)}$, if $\hat{z}_L$ exists.

So, similar as for $U_2(z)$ - see subsection 2.1.7 - $D_2(z)$ can be approximated in the neighborhood of its dominant singularity by:

$$D_2(z) \approx \begin{cases} \dfrac{\hat{K}_2^{(1)}}{\hat{z}_L - z} & \text{if } \hat{z}_L < \hat{z}_B \\ \dfrac{\hat{K}_2^{(2)}}{(\hat{z}_B - z)^{1/2}} & \text{if } \hat{z}_L = \hat{z}_B \\ D_2(\hat{z}_B) - \hat{K}_2^{(3)} (\hat{z}_B - z)^{1/2} & \text{if } \hat{z}_L \text{ does not exist,} \end{cases} \tag{2.127}$$

where the constants $\hat{K}_2^{(i)}$ $(i = 1, 2, 3)$ can be found by investigating $D_2(z)$ (expression (2.104)) in the neighborhood of its dominant singularity. Doing so, we find

$$\hat{K}_2^{(1)} = \frac{(1 - \lambda_T) V(\hat{z}_L) (\hat{z}_L - 1)}{\lambda_2 V'(\hat{z}_L) (A_T'(V(\hat{z}_L)) - 1)} \tag{2.128}$$

$$\hat{K}_2^{(2)} = \frac{(1 - \lambda_T) (\hat{z}_B A_T(V(\hat{z}_B)) - V(\hat{z}_B))}{\lambda_2 K_V (A_T'(V(\hat{z}_B)) - 1)} \tag{2.129}$$

$$\hat{K}_2^{(3)} = \frac{(1 - \lambda_T) K_V (\hat{z}_B - 1) (V(\hat{z}_B) A_T'(V(\hat{z}_B)) - A_T(V(\hat{z}_B)))}{\lambda_2 (V(\hat{z}_B) - A_T(V(\hat{z}_B)))^2}. \tag{2.130}$$

These are calculated in a similar way as the $K_2^{(i)}$ in expressions (2.67), (2.75) and (2.71) respectively.

By using Darboux's theorem on expression (2.127) the tail probabilities of the delay of the class-2 cells are given by

$$d_2(n) \triangleq \text{Prob}[d_2 = n] \tag{2.131}$$

$$\approx \begin{cases} \hat{K}_2^{(1)} \hat{z}_L^{-n-1} & \text{if } \hat{z}_L < \hat{z}_B \\ \dfrac{\hat{K}_2^{(2)} n^{-1/2} \hat{z}_B^{-n}}{\sqrt{\pi \hat{z}_B}} & \text{if } \hat{z}_L = \hat{z}_B \\ \dfrac{\hat{K}_2^{(3)} n^{-3/2} \hat{z}_B^{-n}}{2\sqrt{\pi / \hat{z}_B}} & \text{if } \hat{z}_L \text{ does not exist,} \end{cases} \tag{2.132}$$

with $\hat{K}_2^{(i)}$, $i = 1, 2, 3$, given by (2.128), (2.129) and (2.130) respectively. Again, the first expression shows the typical geometric tail behavior, the third expression the typical non-geometric tail behavior and the second expression the transition type behavior between the two other types.

The probability that a class-2 cell has a delay that exceeds a bound D is then given by

$$\text{Prob}[d_2 > D] \approx \frac{d_2(n)}{\hat{z}_* - 1}, \tag{2.133}$$

with $\hat{z}_*$ the dominant singularity of $D_2(z)$ for D large enough.

Delay random cell

Finally, we calculate the tail probabilities of the delay of a random cell. Its pgf $D(z)$ is given by expression (2.106). The dominant singularity of this function is $\hat{z}_L$ or $\hat{z}_B$, depending on which is smallest.

Note that z_H is also a singularity of $D(z)$, but we will show in this paragraph that $z_H > \hat{z}_B$ and thus that z_H is never dominant. Firstly, since $\hat{z}_B$ satisfies the relation $\hat{z}_B A_1'(V(\hat{z}_B)) = 1$ and since $\hat{z}_B$ is bigger than 1

$$A_1'(V(\hat{z}_B)) < 1. \tag{2.134}$$

Furthermore, since $A_1(z)$ and z intersect in z_H, $A_1'(z_H) > 1$. This combined with (2.134) and the fact that $A_1'(z)$ is assumed to be a strictly increasing function, it follows that

$$V(\hat{z}_B) < z_H. \tag{2.135}$$

Since $V(z) = zA_1(V(z))$ and since $A_1(V(\hat{z}_B)) > 1$ (since $V(\hat{z}_B) > 1$) it follows from the previous inequality that $\hat{z}_B < z_H$, and thus z_H is never a dominant singularity of $D(z)$. It is also intuitively clear that - because of the priority scheduling discipline - the behavior of $d(n)$ for high n will be dominated by the delay of the class-2 cells.

We return to the calculation of the tail probabilities of $D(z)$. From expression (2.105) and the fact that the dominant singularities of $D(z)$ and $D_2(z)$ are equal, it is easily seen that

$$d(n) \triangleq \text{Prob}[d = n] \tag{2.136}$$

$$\approx \frac{\lambda_2}{\lambda_T} d_2(n), \tag{2.137}$$

for large enough n. Since $d_2(n)$ is approximately calculated in expression (2.132), $d(n)$ is approximately determined.

Finally, the probability that the steady-state delay of a random cell is larger than a bound D is given by

$$\text{Prob}[d > D] \approx \frac{\lambda_2}{\lambda_T}\text{Prob}[d_2 > D]. \tag{2.138}$$

2.5 Waiting time

The waiting time of a cell, defined as the number of slots a cell has to wait in the *queue* before getting service, is easily analyzed using the result in the previous section. Indeed since the service times of all cells are equal to one slot, the waiting time of a cell equals the delay of that cell minus 1. Thus, the pgf of the steady-state waiting time of a class-1 cell is given by

$$T_1(z) = \frac{D_1(z)}{z} \tag{2.139}$$

$$= \frac{1-\lambda_1}{\lambda_1}\frac{A_1(z)-1}{z-A_1(z)}. \tag{2.140}$$

Furthermore the pgf of the steady-state waiting time of a class-2 cell yields

$$T_2(z) = \frac{D_2(z)}{z} \tag{2.141}$$

$$= \frac{1-\lambda_T}{\lambda_2}\frac{A_T(V(z))-A_1(V(z))}{V(z)-A_T(V(z))}. \tag{2.142}$$

Finally, the pgf of the steady-state waiting time of a random cell is given by

$$T(z) = \frac{1-\lambda_1}{\lambda_T}\frac{A_1(z)-1}{z-A_1(z)} + \frac{1-\lambda_T}{\lambda_T}\frac{A_T(V(z))-A_1(V(z))}{V(z)-A_T(V(z))}. \tag{2.143}$$

From these pgf's, performance measures can be calculated as is done in the previous section. Similar expressions are found as in the previous section, so we will not go further into detail here.

2.6 Numerical examples

In this section, we will show the influence of the system parameters on the performance measures calculated throughout this chapter. We will specifically focus on the performance measures of the system contents (analyzed in section 2.1) and of the cell delay (analyzed in 2.4).

2.6.1 Input processes

We use the example of the output queueing switch, discussed in section 1.7.2, throughout this section. A conceptual model of such a switch is (again) shown in Figure 2.3. Having (independently distributed) Bernoulli arrivals at the inlets of the switch and having an independent and uniform routing throughout the switch, $A(z_1, z_2)$, the pgf of the number of per-slot class-1 and class-2 arrivals to one of the (output) queues, is given by - see section 1.7.2 -

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2)\right)^N, \tag{2.144}$$

with N the number of inlets and with λ_j the probability that a class-j cell arrives at a randomly chosen inlet. The marginal pgf $A_T(z)$ is then given by

$$A_T(z) = \left(1 - \frac{\lambda_T}{N}(1 - z)\right)^N, \tag{2.145}$$

with $\lambda_T = \lambda_1 + \lambda_2$. The marginal pgf of the number of per-slot arrivals of class-j is given by

$$A_j(z) = \left(1 - \frac{\lambda_j}{N}(1 - z)\right)^N, \tag{2.146}$$

with $j = 1, 2$. The means of the total, class-1 and class-2 number of per-slot arrivals are thus given by λ_T, λ_1 and λ_2 respectively. The variances of these three stochastic variables are given by

$$\text{Var}[a] = \lambda\left(1 - \frac{\lambda}{N}\right), \tag{2.147}$$

with λ equal to λ_T, λ_1 and λ_2 respectively. Finally, the covariance of the numbers of per-slot class-1 and class-2 arrivals is given by

$$\text{Cov}[a_1, a_2] = -\frac{\lambda_1\lambda_2}{N}. \tag{2.148}$$

This covariance is always negative for finite N, which is intuitively clear since having more arrivals of one class in a slot, means that less arrivals of the other class can occur (the maximum number of per-slot cell arrivals is N). If $N \to \infty$, the arrival processes of both classes are uncorrelated (distributed according to Poisson processes), and the covariance thus tends to 0.

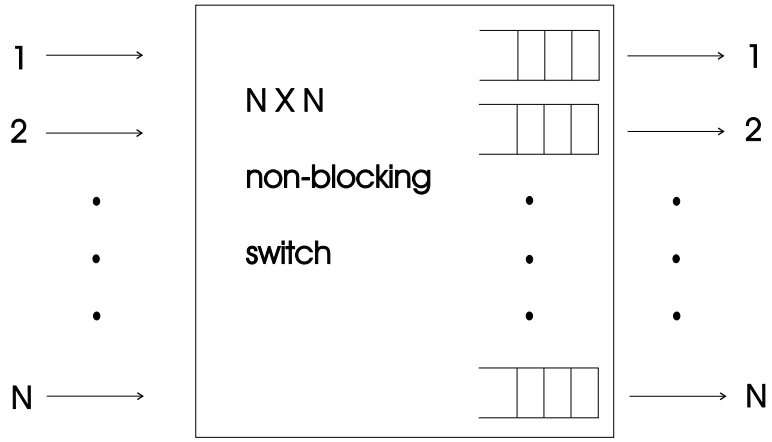


Figure 2.3: An NxN output queueing switch

2.6.2 Influence of load on moments

Firstly, we will show the influence of the load characteristics on the means and variances of the system contents and cell delays. The arrival process is as defined in expression (2.144) with $N = 16$. We define α as the fraction of class-1 arrivals in the overall traffic mix, i.e.,

$$\alpha \triangleq \frac{\lambda_1}{\lambda_T}. \tag{2.149}$$

Mean values and variances of the system contents

In Figures 2.4 and 2.5, the mean values and variances of the system contents of class-1 and class-2 cells are shown as functions of the total arrival rate, when $\alpha = 0.25$, 0.5 and 0.75 respectively. We have also shown the mean value and variance of the system contents of one class (class-1 or class-2) for $\alpha = 0.5$ when a FIFO scheduling discipline is applied. Note that in this case the system contents of class-1 and class-2 are identically distributed. These values can be easily calculated because - in the special case of the arrival process characterized by (2.144) - the joint pgf of the numbers of arrivals of both classes has the feature that it can be written as $A_T(\alpha z_1 + (1-\alpha)z_2)$. I.e., looking at one (arriving) cell, this cell is of class-1 with probability α and of class-2 with probability $1-\alpha$, irrespective of the type of other arrivals. When the scheduling discipline is FIFO, the order of service is random with respect to the type of the cells and thus every cell in the system is also of class-1 with probability α and of class-2 with probability $1-\alpha$. The joint pgf of the system contents of both classes is thus given by $U_T(\alpha z_1 + (1-\alpha)z_2)$, with $U_T(z)$ the pgf of the

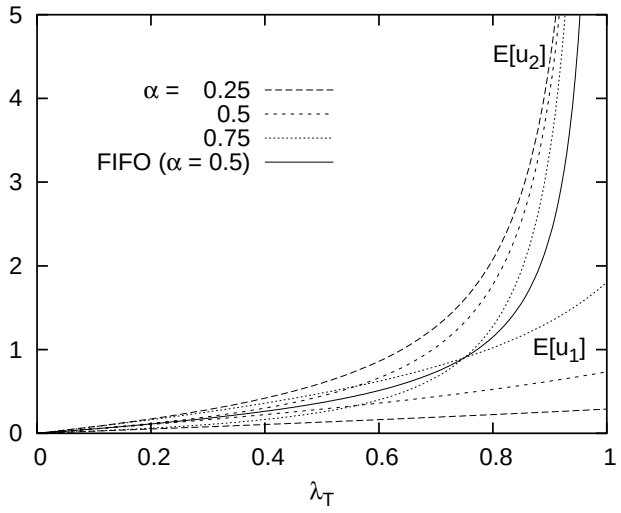


Figure 2.4: Mean value of system contents versus the total arrival rate

total system contents (given by expression (2.18)), and the mean and variance are thus easily obtainable from this pgf. From Figures 2.4 and 2.5, one can see the influence of the priority scheduling discipline: the mean and the variance of the number of class-1 cells in the system are severely reduced by the priority scheduling discipline; the opposite holds for class-2 cells. In addition, it is also clear that the impact of the priority scheduling discipline on the system contents is more important if the (total) load is high. Finally, it also becomes apparent that increasing the fraction of class-1 cells in the overall traffic mix increases the amount of class-1 cells in the system while decreasing the amount of class-2 cells.

Similar conclusions can be drawn from Figures 2.6 and 2.7, which show the mean value and variance respectively of the system contents of both classes versus α for $\lambda_T = 0.3$, 0.6 and 0.9. The mean value and variance of the class-1 system contents increase with the fraction of class-1 cells in the traffic mix, while the opposite holds for the mean value and variance of the class-2 system contents. As can be seen from both figures, the difference between class-1 and class-2 system contents for different values of α can be especially large when the load is high. For high α, the mean and variance of class-1 system contents can be larger than the mean and variance of the class-2 system contents. This is due to the fact that most of the arriving cells are of class-1 for high α and thus the class-1 queue builds up "more" than the class-2 queue (although class-1 cells are served with priority). The most extreme case is when $\alpha = 1$ and thus all cells are of type 1, which means that the class-2 buffer stays empty.

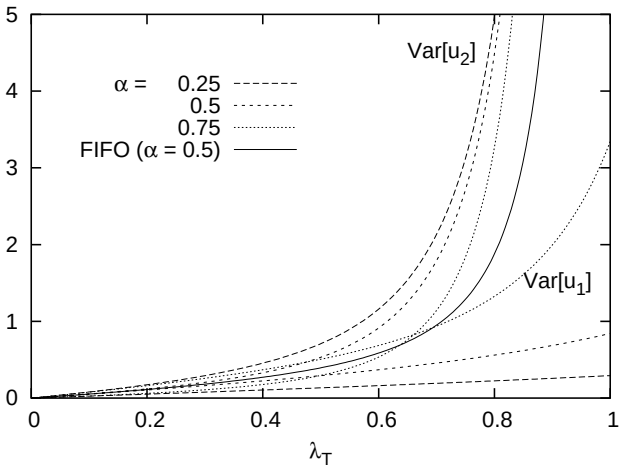


Figure 2.5: Variance of system contents versus the total arrival rate

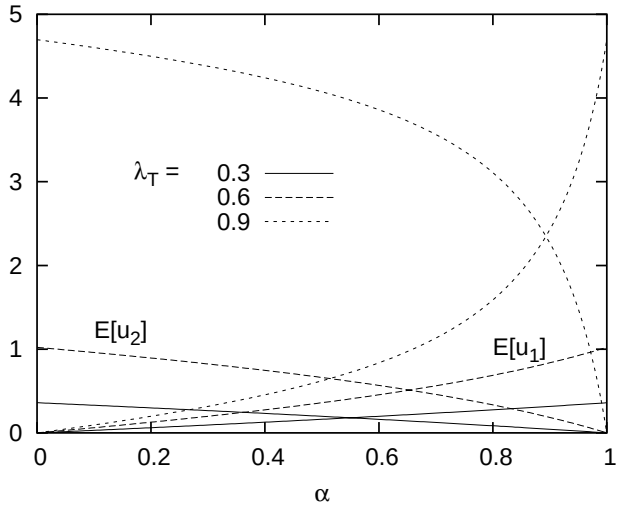


Figure 2.6: Mean value of system contents versus the fraction of class-1 arrivals

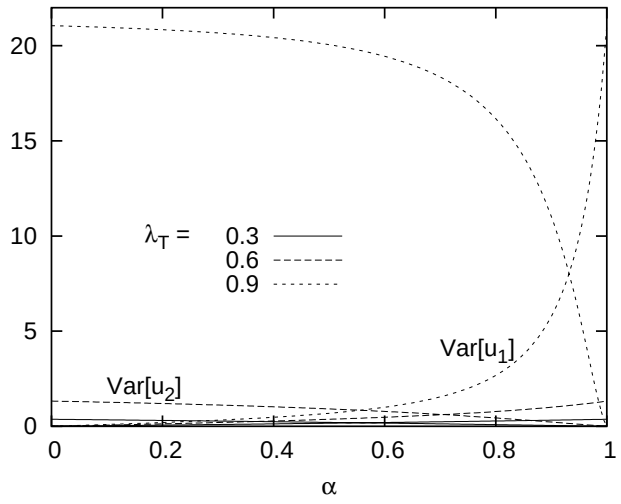


Figure 2.7: Variance of system contents versus the fraction of class-1 arrivals

Correlation coefficient of class-1 and class-2 system contents

In Figure 2.8, the correlation coefficient $\rho_{u_1 u_2}$, which quantifies the correlation between the number of class-1 and class-2 cells in the system at the beginning of a slot, is shown as a function of the total arrival rate for $\alpha = 0.25$, 0.5 and 0.75. We see that $\rho_{u_1 u_2}$ is (slightly) negative when the total load is small, but becomes positive when the total load is large. The reason for this are two counteracting mechanisms. The first one is the switch structure: when more class-1 cells arrive at the switch, there will be less class-2 cells arriving at the same time (since the amount of inlets is limited), and vice versa. This negative correlation between cell arrivals of the two priority classes during a slot shows for small values of λ_T. For these parameter values, there is virtually no queueing and the buffer behavior is mainly determined by the number of arrivals during a single slot. The second mechanism is the priority scheduling discipline. As λ_T (and λ_1) further increases, more and more cells are being queued, and the presence of class-1 cells starts to seriously hinder the transmission of class-2 cells, thereby leading to a positive correlation between u_1 and u_2. Finally, when λ_T approaches 1, the total system contents (and the number of class-2 cells) approaches infinity, due to the system becoming unstable. As a result $\rho_{u_1 u_2}$ approaches 0. We have also shown the correlation coefficient for $\alpha = 0.5$, when a FIFO scheduling discipline is applied. The correlation coefficient is in this case larger than when a priority scheduling discipline is applied (for $\alpha = 0.5$). Since the system contents of both classes go to infinity at the same pace (for $\alpha = 0.5$) when λ_T approaches 1, $\rho_{u_1 u_2}$ approaches 1.

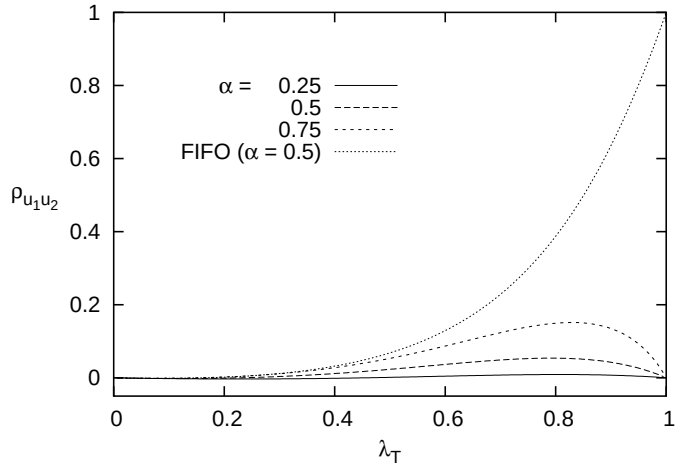


Figure 2.8: Correlation coefficient of system contents versus the total arrival rate

In Figure 2.9, we show the correlation coefficient of the class-1 and class-2 system contents versus the fraction of class-1 cells α, with $\lambda_T = 0.3$, 0.6 and 0.9 respectively. From this figure it is seen that the correlation coefficient increases with increasing α for α not too high (as Figure 2.8 also indicated). This is not true for high α though. For higher α the correlation coefficient decreases with increasing α. This can be understood as follows: since for high α the arriving cells are mainly from class-1, the class-2 queue does not build up dramatically. Again, the extreme case is $\alpha = 1$, i.e., there are no class-2 packets arriving in the queue. For $\alpha = 1$, there is no correlation between the numbers of class-1 and class-2 system contents (the class-2 system contents equals 0, independently of the value of the class-1 system contents).

Mean values and variances of the cell delays

Figures 2.10 and 2.11 show the mean values and the variances of the cell delays of both classes as functions of the total load for $\alpha = 0.25$, 0.5 and 0.75. In order to compare with FIFO scheduling, we have also shown the mean value and variance of the cell delay of any cell in that case. The cell delay is in this particular case the same for class-1 and class-2 cells (independently of α), and can thus be calculated as if there were only one class arriving according to an arrival process with pgf $A(z, z)$. Note that the mean delay in the FIFO case equals the mean delay of a random customer in the system with a priority scheduling discipline. We observe that the influence of a priority scheduling discipline is quite large. The mean delay and the variance of the delay of class-1 cells reduce considerably compared to a queue with a FIFO scheduling discipline. The price to pay is of course a larger mean and variance of the

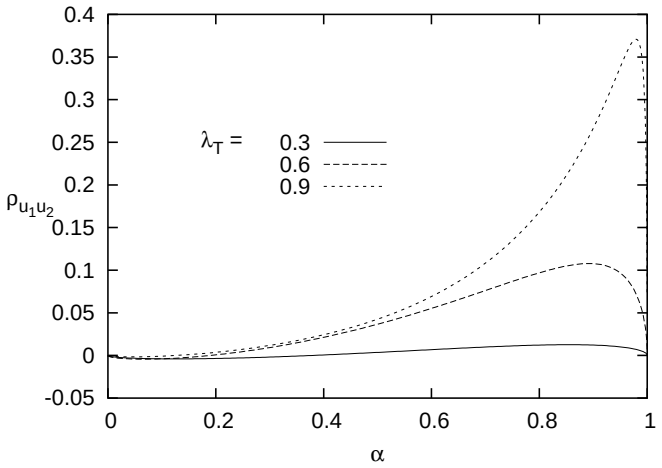


Figure 2.9: Correlation coefficient of system contents versus the fraction of class-1 arrivals

class-2 cell delay. Also note that it follows from these figures that increasing the fraction of class-1 cells in the overall traffic mix, increases the delay characteristics of class-1 and class-2 cells. This is quite obvious: more class-1 cells in the traffic mix means that the class-1 arrival rate increases and thus the delay characteristics of the class-1 cells deteriorates too. Secondly, more class-1 cells in the traffic mix means that more class-1 cells arrive while class-2 cells are waiting in the queue. Since these have priority over the class-2 cells, the delay of the class-2 cells increases as well.

In Figures 2.12 and 2.13, the mean values and variances of the delay of a class-1, class-2 and random cell respectively are shown as functions of the fraction of class-1 cells in the traffic mix, for a total load of 0.6. As expected (and already explained) the mean delay of a random cell is independent of α. The mean delay of the class-1 cells is *always* smaller than the mean delay of a random cell, while the mean delay of a class-2 cell is always larger (except for the 2 boundary cases $\alpha = 0$ and $\alpha = 1$: in these cases we have a single-class system (all cells are of class-1 or of class-2 respectively)). The same is valid for the variances. It is also seen from Figure 2.12 that the mean class-1 and class-2 delay are both increasing functions with α - as already discussed. From Figure 2.13, we conclude that the variance of the delay of a random cell is smallest for $\alpha = 0$ and $\alpha = 1$. In these cases, we have a single-class system with a FIFO scheduling discipline. For other α values, the variance of the delay of a random cell is larger. This is intuitively clear: a priority scheduling discipline is adopted to decrease the delay of a class of cells (class-1) by increasing the delay of the other cells (class-2 cells). By applying the priority scheduling discipline, the variance of the delay of a random cell will increase. As for the

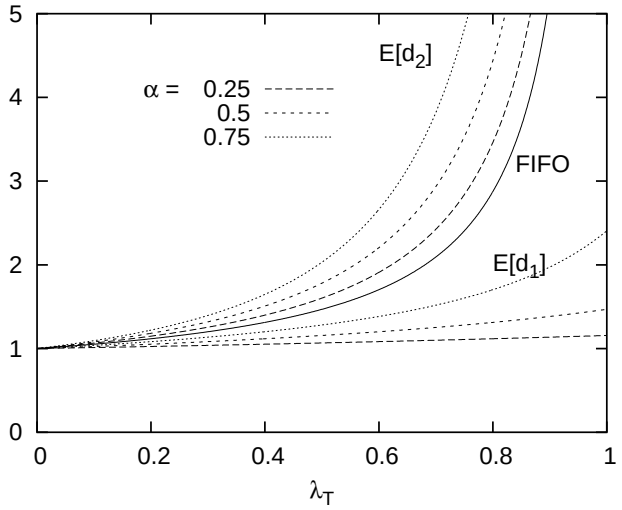


Figure 2.10: Mean value of cell delays versus the total arrival rate

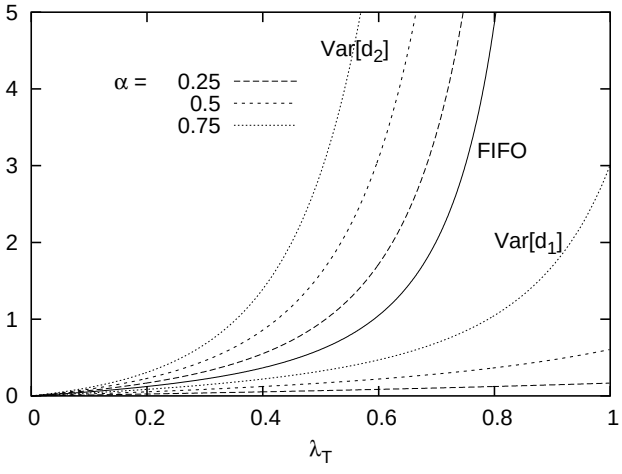


Figure 2.11: Variance of cell delays versus the total arrival rate

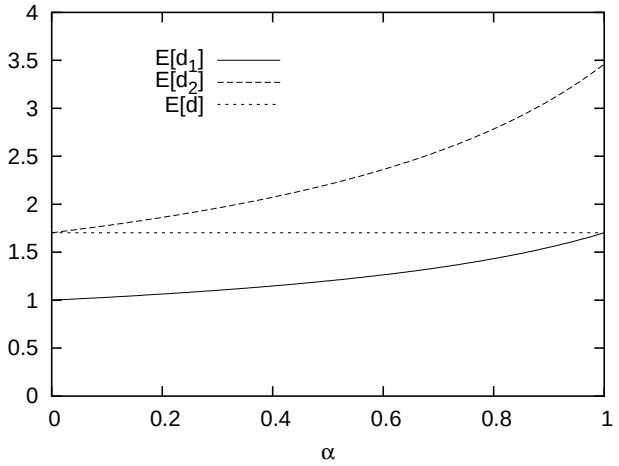


Figure 2.12: Mean cell delays versus the fraction of class-1 cells ($\lambda_T = 0.6$)

mean values, it is seen that the variances of the delay of class-1 and class-2 increase with increasing α.

Concluding this subsection, it is clear that the priority scheduling discipline is especially effective - in reducing the delay of the high-priority cells - when the fraction of these high-priority cells in the traffic mix is (kept) small.

2.6.3 Influence of second order characteristics of the arrival process on mean values

From the expressions of the mean class-1 and class-2 system contents and cell delay, it is easily seen that the second order characteristics of the arrival process (variances and covariances) have an influence on these mean values. For example the mean class-2 system contents and mean class-2 cell delay are seen to be linearly dependent of the variance of the number of per-slot class-1 arrivals, the variance of the number of per-slot class-2 arrivals and the covariance between the number of per-slot class-1 and class-2 arrivals. From expressions (2.147) and (2.148), it can be seen that increasing the parameter N of the (two-dimensional) binomial distribution of the arrival process, Var$[a_1]$, Var$[a_2]$ and Cov$[a_1, a_2]$ are increased, while the arrival rates of both classes are kept constant. In this subsection, we will thus change N in order to study the influence of the second order characteristics of the arrival process.

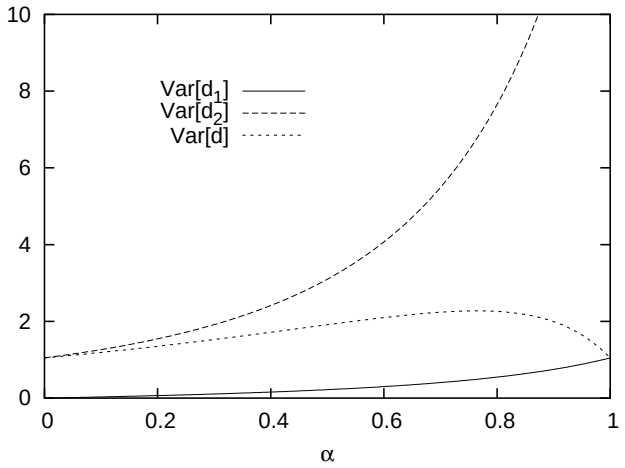


Figure 2.13: Variance of cell delays versus the fraction of class-1 cells ($\lambda_T = 0.6$)

System contents

We show the mean system contents of class-1 and class-2 versus the total arrival rate for $\alpha = 0.25$ and $N = 2$, 4, 16 and ∞ in Figure 2.14. As can be seen from this figure, the second order characteristics play a considerable role in the mean system contents. Especially the mean class-2 system contents increases considerably when N increases. This is because the variance of the number of arrivals of *both* classes and the covariance between these two have impact on the mean class-2 system contents, as opposed to the mean class-1 system contents that is only influenced by the variance of the number of per-slot class-1 arrivals.

Cell delays

In Figure 2.15, the mean class-1 and class-2 cell delay are shown versus the total arrival rate for $\alpha = 0.25$ and $N = 2$, 4, 16 and ∞. Similar conclusions as for the mean system contents can be drawn.

2.6.4 Tail probabilities

We have shown in subsections 2.1.7 and 2.4.6 that the tail probabilities of the class-2 system contents and class-2 cell delay can have three types of behavior, depending on which singularity of $U_2(z)$ or $D_2(z)$ respectively is dominant. In case of the output queueing switch considered in this section (arrival process given by expression (2.144) with $N = 16$), the curves in Figure 2.16 show for

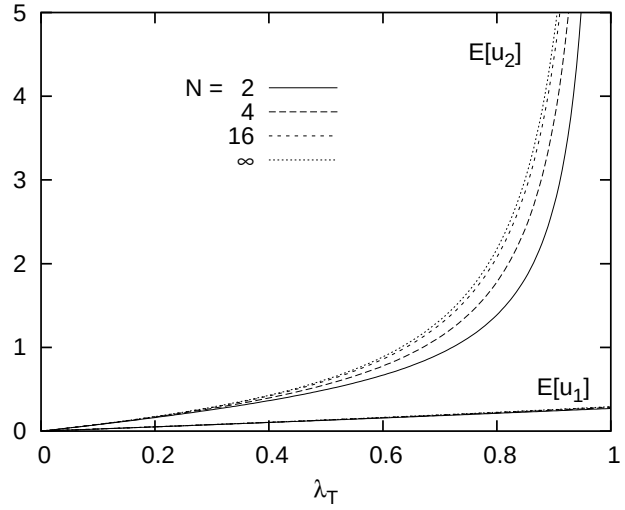


Figure 2.14: Influence of second order characteristics of arrival process on the mean system contents for $\alpha = 0.25$

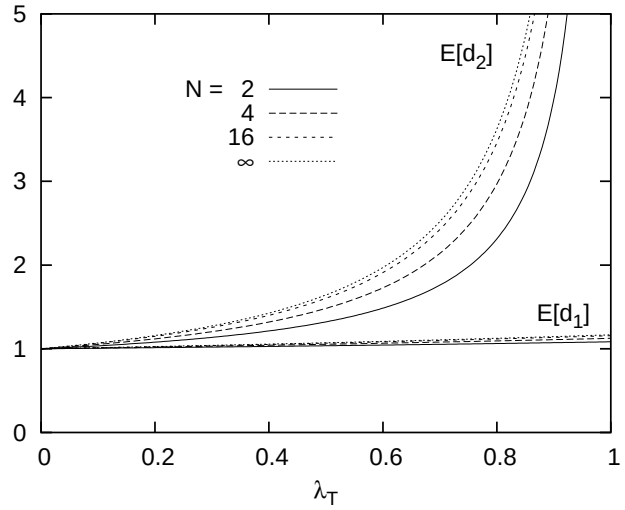


Figure 2.15: Influence of second order characteristics of arrival process on the mean cell delay for $\alpha = 0.25$

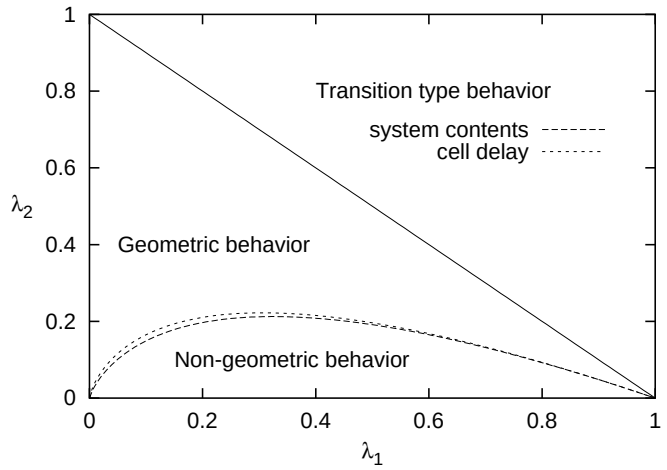


Figure 2.16: Regions for tail behavior as a function of the arrival rates of both classes

which combination of class-1 and class-2 arrival rates the transition type behavior occurs for the system contents and cell delay respectively, i.e., for which combinations of arrival rates the regular pole and the branch point coincide. Above the curves, the tail behavior is geometric, while below the curves the tail behavior is typically non-geometric. E.g. for the system contents curve in Figure 2.16, having an arrival rate combination that is located above (below respectively) the curves means that the regular pole (branch point respectively) of $U_2(z)$ is dominant. For a (λ_1, λ_2)-combination on the curve, these two singularities coincide. Note that in the area above the linear line (defined by $\lambda_1 + \lambda_2 = 1$) in Figure 2.16, the total load is larger than 1, and as a result, the system becomes unstable.

Figures 2.17 and 2.18 show the tail behavior of the system contents and cell delay of class-1 and class-2 cells if $\lambda_1 = 0.4$ and $\lambda_2 = 0.1$ (non-geometric behavior), approximately 0.21 (transition type behavior) and 0.4 (geometric behavior) respectively. The tail behavior of the system contents and cell delay of class-1 cells is of course the same for the three cases, since the arrival process of class-1 cells does not change. We have also compared our approximations with simulation results (marks in the figures). The figures show that the approximations of the class-1, the geometric and transition type tail behavior of system contents and cell delay are very good in these cases. The approximations of the tails of the non-geometric case are not as good, but still satisfactory. Note that the fact that the non-geometric asymptotes are in general not nearly as accurate as the geometric ones is also concluded in Abate and Whitt [1997] - wherein the tail probabilities of the class-2 waiting time in a continuous-time priority queue are approximated.

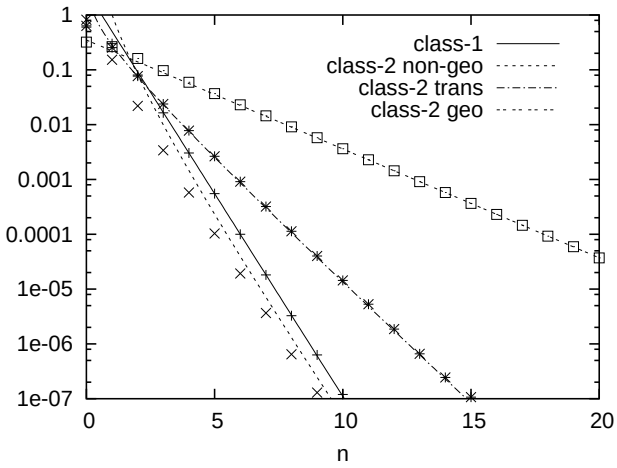


Figure 2.17: Tail behavior of the class-1 and class-2 system contents for some combinations of class-1 and class-2 arrival rates

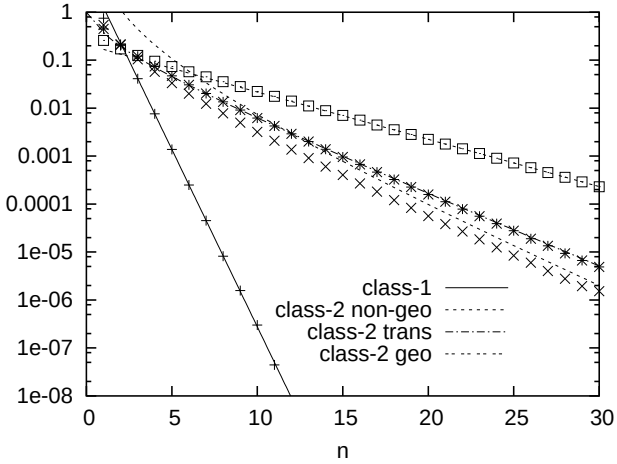


Figure 2.18: Tail behavior of the class-1 and class-2 cell delay for some combinations of class-1 and class-2 arrival rates

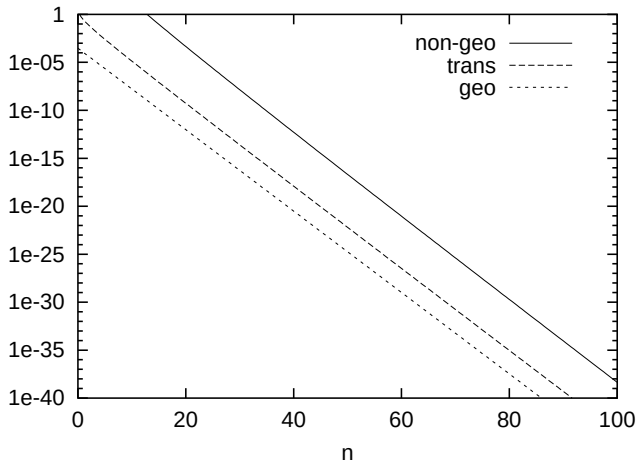


Figure 2.19: Tail behavior of the class-2 system contents near the transition from non-geometrical to geometrical behavior

The approximations of the tails are not good in all cases though, as is illustrated in Figure 2.19. In this figure, the tail probabilities of the class-2 system contents are shown, with the parameters of the transition type behavior of the previous examples ($\lambda_1 = 0.4$ and $\lambda_2 = 0.21$). This can be understood as follows: for $\lambda_1 = 0.4$ constant, we pick three values of λ_2, notably $\lambda_2 = 0.21$ (transition type tail), λ_2 an infinitesimal amount lower than 0.21 (non-geometric) and λ_2 an infinitesimal amount higher than 0.21 (geometric tail). These tail probabilities should be very near to each other, but the figure shows this is not the case. The incorrectness of the geometrical and non-geometrical approximations is due to the single-singularity approximations and/or the approximation of the respective pgf's in their dominant singularity. If both singularities lie near to each other, which is the case near the transition from non-geometric to geometric behavior, the single-singularity approximation is less accurate.

2.7 Concluding remarks

In this chapter, we studied a fairly easy priority queueing system with single-slot service times. We will extend this model to more general service times in the next chapters. The usefulness of this chapter is twofold.

Firstly, we have shown the basic principles of the analysis of a discrete-time priority queue using pgf's and we demonstrated how to calculate the basic performance measures. By using a fairly easy model, the insight in the used

method(s) has not been degraded by too many complexities in the model. Furthermore, the analyses used in the remaining chapters will be highly based on this analysis.

Secondly, although the performance measures of the priority queue in this model are useful in their own right (e.g. in an ATM context), the (pure) influences of the arrival process on the performance measures are easily deductible. This is because the arrival process is the only "stochastic input" in this model. The intuitive explanation of certain behavioral aspects of the priority queue when changing arrival parameters (load, fraction of priority cells in the traffic mix, second order characteristics of the arrival process, ...) can thus more easily be explained.

Chapter 3

Non-preemptive priority

In this chapter, we describe the analysis of a queue with a *non-preemptive* (NP) priority discipline with two priority classes and generally distributed service times. So, whenever the server becomes available, a high-priority (class-1) unit will be scheduled next (if any). If no high-priority traffic is present, a low-priority (class-2) unit is served (if any). The service of a unit cannot be interrupted. Thus when class-1 units arrive while a class-2 unit is in service, the class-1 units have to wait until the class-2 service is completely finished.

For example in telecommunications, the units are not necessarily all of the same length in nowadays multimedia packet-based networks (IP networks for instance). So the model with single-slot service times analyzed in the previous chapter is a too restrictive model to accurately study the queueing phenomena in such networks. Therefore, we will analyze priority queues with general service times in this chapter (and the following chapters). Especially the NP priority discipline has been proposed for packet-based networks, such as the Differentiated Service model for IP networks (see [Xiao and Ni 1999], where traffic of one class, the Premium Service traffic class, has NP priority over all remaining traffic. We will furthermore adopt the packet networks terminology, and call the units *packets* in this chapter. The service time of a packet then equals the number of slots necessary to transmit the packet.

Continuous-time NP priority queues have been introduced by Cobham [1954], according to Miller [1960]. In this latter paper, the pgf of the steady-state system contents and the Laplace transform of the steady-state delay of all priority classes - a general number of priority classes are assumed - are found in case of Poisson arrivals and generally distributed service times. The first two moments of the stochastic variables are calculated from the obtained transform functions. An overview of some other basic (non-)preemptive priority queueing models in continuous-time can be found in the monographs of Kleinrock [1976] and Takagi [1991], and references therein.

Continuous-time NP priority queues with infinite buffer space, *one server* and *no correlation between the arrival processes of the different priority classes* are analyzed in [Marks 1973, Miller 1981, Cidon and Sidi 1990, Takine et al. 1990, Stanford 1991, Takahashi and Miyazawa 1994, Takine et al. 1994a, Sugahara et al. 1995, Takine 1996, 1999, Abate and Whitt 1997, Venkataramani et al. 1997, Boxma et al. 1999, Subramanian and Srikant 2000, Iida et al. 2001, Drekic and Stafford 2002, Karam and Tobagi 2002] and [Isotupa and Stanford 2002]. Marks [1973] gives an algorithm for the calculation of the state probabilities in a NP priority queue with negative exponential interarrival and service times. In [Miller 1981], an NP priority queue with Poisson arrivals and exponential service times is analyzed. The state probabilities of the number of packets of each class in the system are presented in explicit recursive formulas. In [Venkataramani et al. 1997], an NP priority queue with a Markov-modulated Poisson Process (MMPP) as arrival process and with constant service times is studied. This model is applied in an ATM context (hence the constant service times, see chapter 2 for more details). Sugahara et al. [1995] study an NP priority queue with two priority classes, where high-priority packets arrive according to a 2-state Markov Modulated Poisson Process (MMPP) and low-priority packets according to a Poisson process. The service times of all classes are generally distributed and these distributions may differ for different classes. System contents and waiting time are studied using the supplementary variable technique (which we will also use in chapters 4 and 5). Takine et al. [1990] study a polling system and an NP priority system in parallel (with a general number of priority classes/polling stations). The arrival process is a Poisson process and the service times are assumed general. Mean waiting times are obtained. In [Takine et al. 1994a], a NP priority queue with a general number of priority classes, a Markovian Arrival Process (MAP) and general service times (but identically distributed for all priority classes) is studied. More precisely, using matrix-analytic methods and generating function techniques, the mean system contents and mean packet delay are obtained. This analysis is extended to service times with different distributions for the different classes in [Takine 1996] - for two priority classes - and in [Takine 1999] - for a general number of priority classes. Cidon and Sidi [1990] have proposed a recursive computation of the steady-state probabilities of the system contents, using the generating functions of the system contents obtained in [Miller 1960]. In [Takahashi and Miyazawa 1994], a relationship between the distribution of the system contents and the waiting time is obtained for an NP priority queue. Abate and Whitt [1997] quantify the effect of the priority structure on the low-priority steady-state delay tail probabilities (in NP and preemptive resume priority systems). They show that the priority structure tends to make these tail probabilities have a relatively long tail. Boxma et al. [1999] have derived a heavy-traffic limit theorem for the low-priority waiting time when the service times are heavy-tailed. Subramanian and Srikant [2000] calculate the tail probabilities of the low-priority waiting times numerically - from the Laplace-Stieltjes transform - in an NP priority queue with a Markovian Arrival Process (MAP)

and general service times. From their numerical studies, they conclude that these tail probabilities may be non-exponential. Iida et al. [2001], Karam and Tobagi [2002] analyze the delay of the high-priority packets in an NP priority queue with the arrivals broadcasted by a number of multiplexed sources. The interarrival times in the packet stream of each source and the service times of the packets are deterministic. In [Iida et al. 2001], this model is used to model the CBR (Constant Bit Rate) traffic in ATM, while in [Karam and Tobagi 2002] the delay of voice traffic in the Internet is studied. In both papers, the influence of the packet size on the delay is analyzed. In [Drekic and Stafford 2002], a symbolic computation procedure is proposed to calculate (higher order) moments of the system contents and packet delay - starting from the pgf or Laplace-Stieltjes transform of these variables - in (general) priority queues. Isotupa and Stanford [2002] analyze an NP priority queue with Poisson arrivals and phase-type service times. Using matrix-geometric solution techniques, the joint distribution of the number of packets of all classes in the system is derived.

Continuous-time multi-server NP priority queues with a general number of priority classes are analyzed in [Williams 1980, Wagner 1997, Altinkemer et al. 1998, Kao and Wilson 1999, Bose and Pal 2002]. In [Altinkemer et al. 1998, Bose and Pal 2002], a Poisson arrival process and deterministic service times are assumed. In [Altinkemer et al. 1998] the buffer space is infinite while the buffer space is assumed to be finite in [Bose and Pal 2002]. In both papers, approximate results of the mean waiting time of all classes are obtained, using the heavy-traffic assumption. In [Williams 1980, Wagner 1997], the service times are assumed to be identical for all priority classes. These service times are assumed to be exponentially distributed in [Wagner 1997] and generally distributed in [Williams 1980]. In the latter paper the analysis is approximate while it is exact in [Wagner 1997]. Kao and Wilson [1999] study an NP priority queue with Poisson arrivals and exponential service times (which may be different for different classes). Using matrix-analytic methods, several performance measures - such as the mean system contents and waiting times - are derived. The Laplace-Stieltjes transforms of the waiting times of each priority class are calculated.

In most "continuous-time papers" about priority queues, it is assumed that the arrival processes of the different priority classes are all mutually uncorrelated. An *exception* is [Langaris and Katsaros 1995]. Langaris and Katsaros [1995] perform a time-dependent analysis of an NP priority queue with a general number of priority classes and batch arrivals. It is assumed that batches arrive according to a Poisson distribution and that a batch consists of packets from several classes. Laplace-Stieltjes transforms of the busy period of each class and the general busy period as well as the pgf of the system contents are calculated.

Discrete-time priority queues with an NP priority discipline and *without correlation between the arrival processes of the different priority classes* are analyzed in [Rubin and Tsai 1989, Schormans et al. 1991, Choi et al. 1997, Wang et al. 2000]

and [Lee et al. 2003]. In [Rubin and Tsai 1989, Lee et al. 2003], waiting time and delay of a two-class priority queue where the number of arrivals are i.i.d. from slot-to-slot and the service times are generally distributed (and can be different for both classes) are studied using pgf's. In [Schormans et al. 1991], unfinished work and waiting times are analyzed entirely in the probability domain. The arrival process is such that of each class maximum one cell arrives per slot. In [Choi et al. 1997], the NP and preemptive resume priority disciplines are investigated in an ATM packet switch with input buffers and two priority classes. Note that the service times are not deterministic (as one would expect in an ATM context) because a cell at the head of the input queue is not necessarily switched to the desired output directly, since several cells can simultaneously request to be sent to the same output, while only one of them will effectively be transmitted during that slot. The other cells have to wait longer than one slot at the head of their queue - which is incorporated in the service times in the queueing model - leading to service times of more than one slot. A.o., the mean delay of both classes is obtained. Finally, Wang et al. [2000] calculate the mean delay in an NP priority queue and use this in the performance analysis of a multiple-channel slotted ring network with tunable transmitters and fixed receivers.

Discrete-time NP priority queues with *correlation between the arrival processes of different priority classes* are studied in [Takahashi and Hashida 1991] and [Walraevens et al. 2000b,c,d, 2002a, 2003b]. In all these systems the service times are generally distributed and can be different from class-to-class. In [Walraevens et al. 2000b,c,d, 2002a] two-class priority system are analyzed while the number of priority classes is equal to three in [Walraevens et al. 2003b] and assumed general in [Hashida and Takahashi 1991]. In [Takahashi and Hashida 1991], the pgf of the delay of all priority classes is calculated, based on delay-cycles. In [Walraevens et al. 2000b,c,d, 2002a, 2003b], the joint pgf of the system contents of all classes at specific slots is calculated. From this joint pgf, the joint pgf of the system contents of all classes at random slot boundaries (in [Walraevens et al. 2000c,d, 2003b]) and the pgf's of the delays of all priority classes (in [Walraevens et al. 2000b, 2002a, 2003b]) are obtained.

In this chapter, we describe the analysis of a two-class non-preemptive priority buffer with a discrete structured batch arrival process and general class-dependent service times, as in [Walraevens et al. 2000b,c,d, 2002a]. In section 3.2, we calculate the joint pgf of the system contents of both priority classes at specific slot boundaries. All other pgf's and performance measures will be determined, starting from this joint pgf. This procedure is first explained in more detail in section 3.1. The system contents, the queue contents and the unfinished work at the beginning of random slots are analyzed in sections 3.3, 3.4 and 3.5 respectively. Further, the packet delay and waiting time are studied in sections 3.6 and 3.7. Finally, we show the impact of the input parameters on the performance measures through some numerical examples in section 3.8.

3.1 Preliminaries

Since the service times of the packets are generally distributed, the system contents of both classes at the beginning of random slots do not form a Markov chain (as was the case for single-slot service times discussed in chapter 2). Or in other words, the knowledge of the system contents at the beginning of slot k is not enough information to know the system contents at the beginning of slot $k+1$. The solution is to describe the system such that it forms a Markov chain.

There are 2 main directions: the first is defining a number of supplementary stochastic variables at the beginning of all slots, so that the system contents at the beginning of consecutive slots together with these supplementary variables form a Markov chain. This is called the *supplementary variable technique* and is - according to e.g. Chaudhry and Templeton [1983] - first used by Kosten in 1942. The name appears to be due to Cox [1955]. It is for instance used in [Bruneel 1993] for the single-class (i.e. without priorities) version of the buffer studied in this chapter. A second approach is to look at specific slot boundaries instead of at all slot boundaries. These slot boundaries are chosen in such a way that the system contents of both classes at the beginning of these (consecutive) specific slot boundaries form a Markov chain. This Markov chain is called an *embedded Markov chain*, since the Markov chain is embedded in the choice of the chosen slot boundaries.

We will use the latter technique in this section, while we will use the supplementary variable technique in chapters 4 and 5 in the analysis of preemptive priority queues. Note that this supplementary variable technique can also be used in the case of the NP priority queue analyzed in this chapter (an example is found in [Choi et al. 1997]).

3.2 System contents at the beginning of start-slots

We will thus first analyze the system contents at the beginning of so-called *start-slots*. These slots are defined as slots at the beginning of which a service of a packet (if one available) can start. Note that every slot during which the system is empty, is also a start-slot. We denote the system contents of class-j packets at the beginning of the l-th start-slot by $n_{j,l}$ ($j = 1, 2$). In Figure 3.1, a sample of the time axis is shown. Specifically, the location of the start-slots is shown in this figure. The set $\{(n_{1,l}, n_{2,l}), l \geq 1\}$ forms a Markov chain, since the numbers of arrivals of both classes are i.i.d. from slot-to-slot and only random variables during start-slots are involved.

The joint pgf of the $n_{j,l}$, $j = 1, 2$, is denoted by $N_l(z_1, z_2)$, i.e.,

$$N_l(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{n_{1,l}} z_2^{n_{2,l}}\right]. \tag{3.1}$$

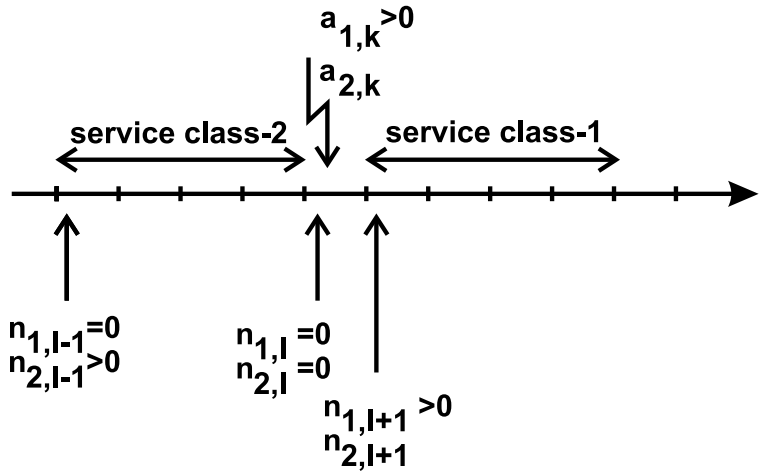


Figure 3.1: Sample of the time-axis in order to show the location of start-slots

Let s_l^* indicate the service time of the packet that enters service at the beginning of start-slot l (which is - by definition - regular slot k) if $n_{1,l}+n_{2,l} > 0$ and equal to 1 if $n_{1,l} = n_{2,l} = 0$. Or alternatively, s_l^* indicates the number of slots between the beginning of the l-th start-slot and the beginning of the $l+1$-th start-slot. The following system equations are then established:

$$n_{1,l+1} = [n_{1,l} - 1]^+ + \sum_{i=0}^{s_l^*-1} a_{1,k+i}; \tag{3.2}$$

$$n_{2,l+1} = \begin{cases} [n_{2,l} - 1]^+ + \sum_{i=0}^{s_l^*-1} a_{2,k+i} & \text{if } n_{1,l} = 0 \\ n_{2,l} + \sum_{i=0}^{s_l^*-1} a_{2,k+i} & \text{if } n_{1,l} > 0 \end{cases}. \tag{3.3}$$

These can be explained as follows: a packet of class-1 is served at the beginning of start-slot l if $n_{1,l} > 0$ and it leaves the system just before start-slot $l+1$. In this case no class-2 packet can be served between start-slots l and $l+1$. A class-2 packet can only be transmitted in this time period if no class-1 packets are present in the system at the beginning of the l-th start-slot, i.e., if $n_{1,l} = 0$. The system contents at the beginning of start-slot $l+1$ are then equal to the number of packets present at the beginning of the previous start-slot minus the one served during the epoch between both start-slots, augmented with the packets that arrive during this same epoch. Notice the similarity between these system equations and the system equations (2.2)-(2.3) of the previous chapter. Indeed, when the service times are equal to one slot (as assumed in the previous chapter), all slots are start-slots and thus $s_l^* = 1$ for all l. Equations (3.2)-(3.3) then simplify to (2.2)-(2.3). The procedure followed in this section to calculate the steady-state joint pgf of the $n_{j,l}$'s will thus also be

similar to the analysis in subsection 2.1.1.

Using system equations (3.2)-(3.3), we derive a relation between $N_l(.,.)$ and $N_{l+1}(.,.)$. Taking into account the statistical independence of the random variables s_l^*, $(n_{1,l}, n_{2,l})$ and $(a_{1,k+i}, a_{2,k+i}), i \geq 0$ respectively, we find - as before, E$[X\{Y\}]$ is defined as E$[X|Y]$Prob$[Y]$ -

$$\begin{aligned} N_{l+1}(z_1,z_2) \triangleq & \mathrm{E}\left[z_1^{n_{1,l+1}} z_2^{n_{2,l+1}}\right] && (3.4)\\ = & \mathrm{E}\left[z_1^{\sum_{i=0}^{s_l^*-1} a_{1,k+i}} z_2^{\sum_{i=0}^{s_l^*-1} a_{2,k+i}} \{n_{1,l}=n_{2,l}=0\}\right] && (3.5)\\ & + \mathrm{E}\left[z_1^{\sum_{i=0}^{s_l^*-1} a_{1,k+i}} z_2^{n_{2,l}+\sum_{i=0}^{s_l^*-1} a_{2,k+i}-1} \{n_{1,l}=0, n_{2,l}>0\}\right] \\ & + \mathrm{E}\left[z_1^{n_{1,l}+\sum_{i=0}^{s_l^*-1} a_{1,k+i}-1} z_2^{n_{2,l}+\sum_{i=0}^{s_l^*-1} a_{2,k+i}} \{n_{1,l}>0\}\right]. \end{aligned}$$

s_l^* is a stochastic variable that equals 1 if $n_{1,l} = n_{2,l} = 0$, that has a (conditional) pgf $S_2(z)$ if $n_{1,l} = 0, n_{2,l} > 0$ and that has a (conditional) pgf $S_1(z)$ if $n_{1,l} > 0$. This leads to

$$\begin{aligned} N_{l+1}(z_1,z_2) = & A(z_1,z_2)\mathrm{Prob}\left[n_{1,l}=n_{2,l}=0\right] && (3.6)\\ & + \frac{S_2(A(z_1,z_2))}{z_2}\mathrm{E}\left[z_2^{n_{2,l}}\{n_{1,l}=0, n_{2,l}>0\}\right] \\ & + \frac{S_1(A(z_1,z_2))}{z_1}\mathrm{E}\left[z_1^{n_{1,l}} z_2^{n_{2,l}}\{n_{1,l}>0\}\right] \\ = & A(z_1,z_2)N_l(0,0) + \frac{S_2(A(z_1,z_2))}{z_2}\left[N_l(0,z_2) - N_l(0,0)\right] && (3.7)\\ & + \frac{S_1(A(z_1,z_2))}{z_1}\left[N_l(z_1,z_2) - N_l(0,z_2)\right], \end{aligned}$$

with $A(z_1, z_2)$ the joint pgf of the numbers of per-slot class-1 and class-2 arrivals and $S_j(z)$ the pgf of the service times of class-j packets ($j = 1, 2$) as defined in chapter 1. In the remainder we will furthermore use the following notation:

$$E_j(z_1,z_2) \triangleq S_j(A(z_1,z_2)), \tag{3.8}$$

with $j = 1, 2$.

We assume that the system is stable. Since the scheduling discipline is work-conserving and since *one* server transmits the packets, this implies that the total arrival load $\rho_T < 1$, with $\rho_T = \lambda_1\mu_1 + \lambda_2\mu_2$ (as a reminder, λ_j is the arrival rate of class-j and μ_j is the mean service time of a class-j packet, $j = 1, 2$). In a stable system, $N_l(z_1, z_2)$ and $N_{l+1}(z_1, z_2)$ converge both to a common steady-

state value for $l \to \infty$:

$$N(z_1, z_2) \triangleq \lim_{l\to\infty} N_l(z_1, z_2). \tag{3.9}$$

By taking the $l \to \infty$ limit of equation (3.7), we obtain:

$$N(z_1, z_2) = \frac{\left\{\begin{array}{l}(z_1E_2(z_1, z_2) - z_2E_1(z_1, z_2))N(0, z_2)\\ +z_1(z_2A(z_1, z_2) - E_2(z_1, z_2))N(0, 0)\end{array}\right\}}{z_2(z_1 - E_1(z_1, z_2))}. \tag{3.10}$$

It now remains for us to determine the unknown function $N(0, z_2)$ and the unknown parameter $N(0, 0)$. This can be done in two steps. First, we notice that $N(z_1, z_2)$ must be bounded for all values of z_1 and z_2 such that $|z_1| < 1$ and $|z_2| < 1$. In particular, this should be true for $z_1 = Y_1(z_2)$, with $Y_1(z_2) \triangleq E_1(Y_1(z_2), z_2)$ and $|z_2| < 1$ (as explained in the appendix). The above implies that if we choose $z_1 = Y_1(z_2)$ in equation (3.10), where $|z_2| < 1$, the denominator of the right-hand side of this equation vanishes. Since $N(z_1, z_2)$ is a pgf, the same must then be true for the numerator, yielding

$$N(0, z_2) = N(0, 0)\frac{z_2A(Y_1(z_2), z_2) - Y_2(z_2)}{z_2 - Y_2(z_2)}, \tag{3.11}$$

with

$$Y_2(z) \triangleq E_2(Y_1(z), z). \tag{3.12}$$

An almost fully determined expression for $N(z_1, z_2)$ can now be derived by combining equations (3.10) and (3.11):

$$N(z_1, z_2) = \frac{N(0, 0)\left\{\begin{array}{l}z_1(z_2A(z_1, z_2) - E_2(z_1, z_2))\\ +Y_2(z_2)(E_1(z_1, z_2) - z_1A(z_1, z_2))\\ +A(Y_1(z_2), z_2)(z_1E_2(z_1, z_2) - z_2E_1(z_1, z_2))\end{array}\right\}}{(z_1 - E_1(z_1, z_2))(z_2 - Y_2(z_2))}. \tag{3.13}$$

Finally, in order to find an expression for $N(0, 0)$, we put $z_1 = z_2 = 1$ and use de l'Hôpital's rule in equation (3.13). Therefore, we need the first derivative of $Y_1(z)$ (at $z = 1$). $Y_1'(z)$ is given by

$$Y_1'(z) = E_1^{(1)}(Y_1(z), z)Y_1'(z) + E_1^{(2)}(Y_1(z), z) \tag{3.14}$$

$$= \frac{E_1^{(2)}(Y_1(z), z)}{1 - E_1^{(1)}(Y_1(z), z)}, \tag{3.15}$$

with $E_1^{(j)}(x,y) = \left.\frac{\partial E_1(z_1,z_2)}{\partial z_j}\right|_{z_1=x,z_2=y}$. Using the definition of $E_1(z_1,z_2)$ and the fact that $Y_1(1) = 1$ - we will prove later on that $Y_1(z)$ is a pgf - it follows that

$$Y_1'(1) = \frac{\lambda_2\mu_1}{1-\rho_1}, \tag{3.16}$$

with $\rho_1 = \lambda_1\mu_1$, the class-1 arrival load. We then obtain $N(0,0)$:

$$N(0,0) = \frac{1-\rho_T}{1-\rho_T+\lambda_T}, \tag{3.17}$$

with λ_T the total arrival rate. Substituting this expression in (3.13) finalizes the calculation of $N(z_1,z_2)$.

We will use expression (3.13) as a starting-point for the analysis of the system contents at random slot boundaries and for the analysis of all other stochastic variables studied in this chapter.

Note that expression (3.13) equals expression (2.13) when the service times are deterministically equal to 1 slot (i.e., when $E_j(z_1,z_2) = A(z_1,z_2)$, $j = 1,2$), as expected.

3.3 System contents at the beginning of random slots

In this section, we analyze the system contents at the beginning of arbitrary slots. First, we will determine the joint pgf of the steady-state system contents of both classes. From this pgf, marginal pgf's and performance measures are calculated.

3.3.1 Calculation of the joint pgf $U(z_1,z_2)$

Denoting the class-j system contents at the beginning of slot k as $u_{j,k}$, the joint pgf of the system contents of both priority classes at the beginning of slot k is defined as:

$$U_k(z_1,z_2) \triangleq \mathrm{E}[z_1^{u_{1,k}} z_2^{u_{2,k}}]. \tag{3.18}$$

In order to derive an expression for $U_k(z_1,z_2)$, we have to know the status of the server during slot k. There are 3 possibilities: the server can be idle, a

class-2 packet or a class-1 packet can be in service during slot k. This yields

$$\begin{aligned} U_k(z_1,z_2) =& \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\{\text{no service}\}\right] + \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\{\text{service class-2}\}\right] \\ &+ \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\{\text{service class-1}\}\right], \end{aligned} \tag{3.19}$$

with "no service" and "service class-j" ($j = 1, 2$) short for the events that there is no service during slot k and that a class-j packet is being served during slot k respectively. If slot k is a start-slot, we will assume that it is start-slot l. If slot k is not a start-slot on the other hand, the last start-slot preceding slot k is denoted as start-slot l. Equation (3.19) then becomes

$$\begin{aligned} U_k(z_1,z_2) =& \text{Prob[no service]}\mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}} | n_{1,l} = n_{2,l} = 0\right] \\ &+ \text{Prob[service class-2]}\mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}} | n_{1,l} = 0, n_{2,l} > 0\right] \\ &+ \text{Prob[service class-1]}\mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}} | n_{1,l} > 0\right]. \end{aligned} \tag{3.20}$$

This can be understood as follows: the server is idle during slot k if there were no packets in the system at the beginning of start-slot l, a class-2 packet is being served during slot k if there were no class-1 packets and at least one class-2 packet in the system at the beginning of start-slot l and a class-1 packet is in service during slot k if there was at least one class-1 packet in the system at the beginning of start-slot l. Thus the three conditions "no service", "service class-2 packet" and "service class-1 packet" imply "$n_{1,l} = n_{2,l} = 0$", "$n_{1,l} = 0, n_{2,l} > 0$" and "$n_{1,l} > 0$" respectively. We denote the elapsed service time of the packet in service (if any) during slot k by s_k^+. This is thus the number of slots between the beginning of start-slot l and the beginning of slot k. The system contents at the beginning of slot k is a superposition of the system contents at the beginning of start-slot l and the arrivals during s_k^+, yielding

$$\begin{aligned} U_k(z_1,z_2) =& \text{Prob[no service]} + \text{Prob[service class-2]} \\ &\times \mathrm{E}\left[z_1^{\sum_{i=1}^{s_k^+} a_{1,k-i}} z_2^{n_{2,l}+\sum_{i=1}^{s_k^+} a_{2,k-i}} | n_{1,l} = 0, n_{2,l} > 0\right] \\ &+ \text{Prob[service class-1]} \\ &\times \mathrm{E}\left[z_1^{n_{1,l}+\sum_{i=1}^{s_k^+} a_{1,k-i}} z_2^{n_{2,l}+\sum_{i=1}^{s_k^+} a_{2,k-i}} | n_{1,l} > 0\right] \end{aligned} \tag{3.21}$$

$$\begin{aligned} =& \text{Prob[no service]} \\ &+ \text{Prob[service class-2]} S_{2,k}^+(A(z_1,z_2)) \frac{N_l(0,z_2) - N_l(0,0)}{N_l(0,1) - N_l(0,0)} \\ &+ \text{Prob[service class-1]} S_{1,k}^+(A(z_1,z_2)) \frac{N_l(z_1,z_2) - N_l(0,z_2)}{1 - N_l(0,1)}, \end{aligned} \tag{3.22}$$

with $N_l(z_1, z_2)$ the joint pgf of the system contents of both classes at the beginning of start-slot l (as defined in the previous section) and $S^+_{j,k}(z)$ ($j = 1, 2$) defined as the pgf of the elapsed service time of the class-j packet in service at the beginning of slot k.

We denote the steady-state version of $U_k(z_1, z_2)$ by $U(z_1, z_2)$, i.e.,

$$U(z_1, z_2) \triangleq \lim_{k\to\infty} U_k(z_1, z_2). \tag{3.23}$$

We first determine the probabilities of the events "no service", "service class-1" and "service class-2" in steady-state, i.e., for a random slot in steady-state. These are given by

$$\text{Prob[no service]} = U(0,0) \tag{3.24}$$

$$\text{Prob[service class-1]} = (1 - U(0,0))\frac{\rho_1}{\rho_T} \tag{3.25}$$

$$\text{Prob[service class-2]} = (1 - U(0,0))\frac{\rho_2}{\rho_T}. \tag{3.26}$$

This is found as follows: the server is idle during a random slot iff the system was empty at the beginning of that slot. This leads to the first expression. On the other hand, if the server is busy during slot k (with probability $1-U(0,0)$), a class-j packet is being served with probability ρ_j/ρ_T ($j = 1, 2$). Substituting expressions (3.24)-(3.26) in the steady-state version of expression (3.22) (thus for $k \to \infty$), we find

$$\begin{aligned} U(z_1, z_2) =& U(0,0) + \frac{1 - U(0,0)}{\rho_T}\left\{\rho_2 S^+_2(A(z_1, z_2))\frac{N(0, z_2) - N(0,0)}{N(0,1) - N(0,0)}\right. \\ &\left. +\rho_1 S^+_1(A(z_1, z_2))\frac{N(z_1, z_2) - N(0, z_2)}{1 - N(0,1)}\right\}, \end{aligned} \tag{3.27}$$

with $N(z_1, z_2)$ and $S^+_j(z)$ ($j = 1, 2$) the steady-state versions of $N_l(z_1, z_2)$ and $S^+_{j,k}(z)$ respectively. Secondly, it is shown in e.g. Bruneel and Kim [1993] that $S^+_j(z)$ yields - keeping in mind that choosing a slot in a larger service time has a higher probability -

$$S^+_j(z) = \frac{S_j(z) - 1}{\mu_j(z - 1)}, \tag{3.28}$$

for $j = 1, 2$. We will show the calculations of this pgf - in a more general setting - further in this chapter. Substituting this expression in (3.27), we find

$$\begin{aligned} U(z_1, z_2) =& U(0,0) + \frac{1 - U(0,0)}{\rho_T(A(z_1, z_2) - 1)}\left\{\lambda_2(E_2(z_1, z_2) - 1)\frac{N(0, z_2) - N(0,0)}{N(0,1) - N(0,0)}\right. \end{aligned} \tag{3.29}$$

$$+\lambda_1(E_1(z_1,z_2)-1)\frac{N(z_1,z_2)-N(0,z_2)}{1-N(0,1)}\Bigg\}.$$

Finally, we calculate $U(0,0)$. Keeping in mind that if the server is idle during slot k, slot k is a start-slot, $U(0,0)$ is found as follows:

$$\begin{aligned}U(0,0) &= \lim_{k\to\infty} \text{Prob}[u_{1,k}=u_{2,k}=0] && (3.30)\\ &= \lim_{k,l\to\infty} \text{Prob}[n_{1,l}=n_{2,l}=0 \text{ and slot } k \text{ is a start-slot}] && (3.31)\\ &= \lim_{k,l\to\infty} \text{Prob}[n_{1,l}=n_{2,l}=0| \text{ slot } k \text{ is a start-slot}] && (3.32)\\ &\quad \times \text{Prob}[\text{slot } k \text{ is a start-slot}]\end{aligned}$$

There are three possibilities for slot k to be a start-slot: the system is empty at the beginning of slot k, slot k is the first slot of the service time of a class-1 packet or slot k is the first slot of the service time of a class-2 packet. $U(0,0)$ then becomes

$$\begin{aligned}U(0,0) =& N(0,0)\left[U(0,0)+\frac{1-U(0,0)}{\mu_1}\frac{\rho_1}{\rho_T}+\frac{1-U(0,0)}{\mu_2}\frac{\rho_2}{\rho_T}\right] && (3.33)\\ =& 1-\rho_T, && (3.34)\end{aligned}$$

where we used expression (3.17) for $N(0,0)$. Finally, using equations (3.13) and (3.34) in equation (3.29), we derive a fully determined version for $U(z_1,z_2)$:

$$\begin{aligned}U(z_1,z_2) =& (1-\rho_T)\frac{E_1(z_1,z_2)(z_1-1)}{z_1-E_1(z_1,z_2)}+(1-\rho_T) && (3.35)\\ &\times \frac{(A(Y_1(z_2),z_2)-1)\left\{\begin{array}{l}z_1E_2(z_1,z_2)(E_1(z_1,z_2)-1)\\ +z_2E_1(z_1,z_2)(1-E_2(z_1,z_2))\\ +z_1z_2(E_2(z_1,z_2)-E_1(z_1,z_2))\end{array}\right\}}{(A(z_1,z_2)-1)(z_1-E_1(z_1,z_2))(z_2-Y_2(z_2))}.\end{aligned}$$

From this two-dimensional pgf, we calculate some marginal pgf's in the following subsections.

Note that for the special case of deterministic service times of one slot, expression (3.35) equals expression (2.13), as expected.

3.3.2 The marginal pgf $U_T(z)$

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

$$U_T(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{T,k}}\right] \qquad (3.36)$$

$$=U(z,z) \tag{3.37}$$

$$=(1-\rho_T)\left[\frac{S_1(A_T(z))(z-1)}{z-S_1(A_T(z))}+\frac{A(Y_1(z),z)-1}{A_T(z)-1}\right. \tag{3.38}$$
$$\left.\times\frac{z(z-1)(S_2(A_T(z))-S_1(A_T(z)))}{(z-S_1(A_T(z)))(z-Y_2(z))}\right].$$

In the special case that $S_1(z) = S_2(z)(\equiv S(z))$, i.e., when the distributions of the service times of class-1 and class-2 priority packets are the same, $U_T(z)$ becomes

$$U_T(z) =(1-\rho_T)\frac{S(A_T(z))(z-1)}{z-S(A_T(z))}. \tag{3.39}$$

This is the expression of the pgf of the system contents in a single-class queue with $A_T(z)$ the pgf of the number of per-slot arrivals, $S(z)$ the pgf of the service times and with a FIFO scheduling discipline (see e.g. Bruneel [1993]). Indeed, if all packets have the same service distribution, the scheduling discipline does not influence the total system contents, as long as the scheduling discipline is work-conserving, independent of the actual service times, and as long as service of packets cannot be preempted. It is also intuitively clear that when $S_1(z) \neq S_2(z)$, the (total) system contents are not equally distributed when applying a priority or a FIFO scheduling discipline.

3.3.3 The marginal pgf $U_1(z)$

From the two-dimensional pgf $U(z_1, z_2)$, we derive an expression for the pgf of the system contents of class-1 packets at the beginning of an arbitrary slot - denoted by $U_1(z)$ - yielding

$$U_1(z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[z^{u_{1,k}}\right] \tag{3.40}$$

$$=U(z,1) \tag{3.41}$$

$$=(1-\rho_1)\frac{S_1(A_1(z))(z-1)}{z-S_1(A_1(z))}\left[\frac{1-\rho_T}{1-\rho_1}+\frac{\rho_2}{1-\rho_1}\frac{S_2(A_1(z))-1}{\mu_2(A_1(z)-1)}\right], \tag{3.42}$$

which is found from expression (3.35) and using de l'Hôpital's rule.

The class-1 contents can be analyzed in an alternative way, namely by looking at the class-1 system as a single-class system with (multiple) *vacations*. Indeed, for class-1 packets it is as if the server goes on vacation when the class-1 system gets empty. During this vacation, a class-2 packet is served, if any, or the vacation only lasts for a slot when no class-2 packets are present in the system at that time instant. Coming back from a vacation, the server starts serving the class-1 system again, or takes another vacation when the class-1 system is still empty, and so on. In the remainder of this subsection, we will call a vacation

when the original system is empty a *vacation of type-I* and a vacation when a class-2 packet of the original priority system is served a *vacation of type-II*. Thus a vacation of type-I lasts one slot, while the length of a vacation of type-II has the same distribution as a class-2 service time. Since our "constructed" vacation queue is part of a broader class of *generalized vacation queues*, we can use the *stochastic decomposition property* of vacation queues (see e.g. Takagi [1991] and Fiems [2004]). In the context of our system, this property states that the pgf $U_1(z)$ of the system contents of the vacation system equals the product of the pgf $U_1^{(1)}(z)$ of the system contents of the system *without* vacations and the pgf $U_1^{(2)}(z)$ of the system contents at the beginning of a random vacation slot. The first pgf is given by - see e.g. Bruneel [1993] -

$$U_1^{(1)}(z) = (1-\rho_1)\frac{S_1(A_1(z))(z-1)}{z - S_1(A_1(z))}. \tag{3.43}$$

The second pgf is calculated as follows: first a random vacation slot is tagged. Since the system is in vacation iff the original priority system is not serving a class-1 packet, the probability that the tagged vacation slot is a slot in a vacation of type-I equals:

$$\begin{aligned}
\text{Prob[vacation slot type-I]} &= \text{Prob[no service}|\text{no service or service class-2]} && (3.44)\\
&= \frac{\text{Prob[no service]}}{\text{Prob[no service or service class-2]}} && (3.45)\\
&= \frac{1-\rho_T}{1-\rho_1}, && (3.46)
\end{aligned}$$

and thus

$$\begin{aligned}
\text{Prob[vacation slot type-II]} &= 1 - \frac{1-\rho_T}{1-\rho_1} && (3.47)\\
&= \frac{\rho_2}{1-\rho_1}. && (3.48)
\end{aligned}$$

When a vacation is of type-I, it lasts one slot and the vacation system is empty at the beginning of that slot (the server of the original system is idle during this slot). When a vacation slot is of type-II, the original system is serving a class-2 packet. Since the vacation system was empty at the beginning of the vacation which the tagged vacation slot is part of, the system contents at the beginning of the tagged vacation slot then equals the number of arrivals during the elapsed part of the vacation (or - translated to the original system - the number of class-1 packet arrivals during the elapsed part of a class-2

service time). $U_1^{(2)}(z)$ thus equals

$$\begin{aligned} U_1^{(2)}(z) =& \text{Prob[vacation slot type-I]} && (3.49)\\ &+ \text{Prob[vacation slot type-II]}\frac{S_2(A_1(z))-1}{\mu_2(A_1(z)-1)} \\ =& \frac{1-\rho_T}{1-\rho_1} + \frac{\rho_2}{1-\rho_1}\frac{S_2(A_1(z))-1}{\mu_2(A_1(z)-1)}. && (3.50) \end{aligned}$$

Using expressions (3.43) and (3.50) in the stochastic decomposition property

$$U_1(z) = U_1^{(1)}(z)U_1^{(2)}(z), \tag{3.51}$$

then indeed yields expression (3.42) for $U_1(z)$.

3.3.4 The marginal pgf $U_2(z)$

The pgf of the class-2 system contents is also calculated from $U(z_1, z_2)$ (expression (3.35)), yielding

$$\begin{aligned} U_2(z) \triangleq& \lim_{k\to\infty} \mathrm{E}\left[z^{u_{2,k}}\right] && (3.52)\\ =& U(1,z) && (3.53)\\ =& (1-\rho_T)\frac{S_2(A_2(z))(z-1)}{z-Y_2(z)}\frac{A(Y_1(z),z)-1}{A_2(z)-1}. && (3.54) \end{aligned}$$

3.3.5 Calculation of moments

In this subsection, we give the expressions of the mean total, class-1 and class-2 system contents respectively.

The mean total system contents is found by taking the first derivative of expression (3.38) and substituting z by 1:

$$\begin{aligned} \mathrm{E}[u_T] =& U_T'(1) && (3.55)\\ =& \frac{\rho_T}{2} + \frac{\mu_1 \mathrm{Var}[a_T]}{2(1-\rho_T)} - \frac{\mu_1\lambda_2(\mu_2-\mu_1)\mathrm{Var}[a_1]}{2(1-\rho_T)(1-\rho_1)} + \frac{(\mu_2-\mu_1)\mathrm{Var}[a_2]}{2(1-\rho_T)} && (3.56)\\ &+ \frac{(\lambda_1\mathrm{Var}[s_1]+\lambda_2\mathrm{Var}[s_2])(\lambda_1(1-\rho_1)+\lambda_2(1-\mu_2\lambda_1))}{2(1-\rho_T)(1-\rho_1)} \\ &+ \frac{\lambda_1\lambda_2(\mu_2-1)(\mu_2-\mu_1)}{2(1-\rho_1)}. \end{aligned}$$

Note that this expression simplifies considerably in the special case that the service times of both classes are equally distributed, or more specifically - in

the case of the mean total system contents - when $\mu_1 = \mu_2 (= \mu)$ and $\text{Var}[s_1] = \text{Var}[s_2] (= \text{Var}[s])$, yielding

$$\text{E}[u_T] = \frac{\rho_T}{2} + \frac{\mu \text{Var}[a_T]}{2(1-\rho_T)} + \frac{\lambda_T^2 \text{Var}[s]}{2(1-\rho_T)}. \tag{3.57}$$

As expected (see also the discussion in subsection 3.3.2), this is the same expression as for the mean system contents in a single-class queue with a FIFO scheduling discipline with a_T the number of arrivals in a random slot (with mean λ_T) and s the service time of a random packet (with mean μ).

Furthermore, the mean class-1 system contents is found by taking the first derivative of $U_1(z)$ and substituting z by 1:

$$\text{E}[u_1] = U_1'(1) \tag{3.58}$$

$$= \frac{\rho_1}{2} + \frac{\mu_1 \text{Var}[a_1]}{2(1-\rho_1)} + \frac{\lambda_1^2 \text{Var}[s_1]}{2(1-\rho_1)} + \frac{\lambda_1 \lambda_2 (\mu_2(\mu_2 - 1) + \text{Var}[s_2])}{2(1-\rho_1)}. \tag{3.59}$$

The sum of the first three terms of this expression equals the mean system contents in a single-class system with only class-1 arrivals (see also the resemblance with expression (3.57)) while the last term of this expression represents the influence of the class-2 traffic on the mean class-1 contents. It is seen that the class-2 arrival rate and the mean and variance of the service times of class-2 packets influence the mean class-1 system contents, while the variance of the number of per-slot class-2 arrivals does not.

Finally, the mean class-2 system contents is given by

$$\text{E}[u_2] = U_2'(1) \tag{3.60}$$

$$= \frac{\rho_2}{2} + \frac{\mu_1^2 \lambda_2 \text{Var}[a_1]}{2(1-\rho_T)(1-\rho_1)} + \frac{\mu_2 \text{Var}[a_2]}{2(1-\rho_T)} + \frac{\mu_1 \text{Cov}[a_1, a_2]}{1-\rho_T} + \frac{\lambda_2(\lambda_1 \text{Var}[s_1] + \lambda_2 \text{Var}[s_2])}{2(1-\rho_T)(1-\rho_1)} - \frac{\rho_1 \lambda_2 (\mu_2 - 1)}{2(1-\rho_1)}. \tag{3.61}$$

It is easily verified that equations (3.56), (3.59) and (3.61) satisfy $\text{E}[u_T] = \text{E}[u_1] + \text{E}[u_2]$.

In a similar way, expressions for the variance (and higher moments) of the system contents can be calculated by taking the appropriate derivatives of the respective pgf's. As in chapter 2, we will not show the expressions of the higher moments, but we will show figures of second order moments in the numerical examples later on.

3.3.6 Calculation of tail probabilities

In this subsection, we show (approximate) expressions of the tail probabilities of the total, class-1 and class-2 system contents, i.e., we approximate Prob$[u_T = n]$, Prob$[u_1 = n]$ and Prob$[u_2 = n]$ for high enough n. We will not go into (much) detail about the calculations, but we refer to the corresponding subsection in the previous chapter (subsection 2.1.7).

Note that we assume for the reasoning in the remainder that the pgf's of the arrival and service processes ($A_T(z)$, the $A_j(z)$ and the $S_j(z)$) and their derivatives go to infinity for z equal to their radii of convergence or for $z \to \infty$ (which is correct for most 'normally' applied arrival distributions). For most pgf's that do not fulfil these assumptions, the reasoning in this subsection can be adjusted, but this is not the main topic of this dissertation (see also section 2.1.7).

Since in the expressions of the pgf's $U_T(z)$ and $U_2(z)$ the function $Y_1(z)$ appears, we will first show the behavior of this function in its dominant singularity.

Behavior of $Y_1(z)$ and $Y_2(z)$ in the neighborhood of the dominant branch-point

$Y_1(z)$ is implicitly defined as $Y_1(z) = E_1(Y_1(z), z)$. As discussed in subsection 2.1.7, this function has a dominant branch-point singularity z_B where $Y_1'(z)$ becomes infinite, i.e.,

$$E_1^{(1)}(Y_1(z_B), z_B) = 1, \tag{3.62}$$

with $E_1^{(j)}(x, y) = \left.\dfrac{\partial E_1(z_1, z_2)}{\partial z_j}\right|_{z_1=x, z_2=y}$, $j = 1, 2$. In the neighborhood of this (dominant) singularity, $Y_1(z)$ is approximately given by

$$Y_1(z) \approx Y_1(z_B) - K_{Y_1} (z_B - z)^{1/2}, \tag{3.63}$$

with

$$K_{Y_1} = \sqrt{\frac{2E_1^{(2)}(Y_1(z_B), z_B)}{E_1^{(11)}(Y_1(z_B), z_B)}}, \tag{3.64}$$

which is found by substituting $z = z_B$ in expression (3.63) and using the definition of $Y_1(z)$. Using the definition $E_1(z_1, z_2) = S_1(A(z_1, z_2))$, this formula is

transformed into

$$K_{Y_1} = \sqrt{\frac{2A^{(2)}(Y_1(z_B), z_B)}{A^{(11)}(Y_1(z_B), z_B) + S_1''(A(Y_1(z_B), z_B)) \left(A^{(1)}(Y_1(z_B), z_B)\right)^3}}. \quad (3.65)$$

z_B is also a branch-point of $Y_2(z)$, since

$$Y_2'(z) = S_2'(A(Y_1(z), z)) \left[A^{(1)}(Y_1(z), z)Y_1'(z) + A^{(2)}(Y_1(z), z)\right]. \quad (3.66)$$

Thus in the neighborhood of z_B, we find

$$Y_2(z) \approx Y_2(z_B) - K_{Y_2}\sqrt{z_B - z}, \quad (3.67)$$

with

$$K_{Y_2} = K_{Y_1} S_2'(A(Y_1(z_B), z_B))A^{(1)}(Y_1(z_B), z_B). \quad (3.68)$$

Total system contents

We now concentrate first on the total system contents. In general, $Y_1(z)$ appears in the expression of $U_T(z)$. In the special case of $S_1(z) = S_2(z)$ for all z (i.e., when $S_1(.)$ and $S_2(.)$ are identical functions), this is not the case though. We will first concentrate on this special case and later on discuss the tail probabilities of u_T when $S_1(z) \neq S_2(z)$.

When $S_1(z) = S_2(z)$, we have expression (3.39) for $U_T(z)$. The dominant singularity of this expression is the zero on the positive real axis (> 1) of $z - S_1(A_T(z))$, denoted as z_T, and this is an ordinary pole with multiplicity 1. In the neighborhood of this pole, $U_T(z)$ is thus approximated by

$$U_T(z) \approx \frac{K_T}{z_T - z}, \quad (3.69)$$

with

$$K_T = \frac{(1 - \rho_T)z_T(z_T - 1)}{S_1'(A_T(z_T))A_T'(z_T) - 1}. \quad (3.70)$$

Using Darboux's theorem, the tail probabilities of u_T are thus given by

$$u_T(n) \triangleq \text{Prob}[u_T = n] \quad (3.71)$$

$$= K_T z_T^{-n-1}, \quad (3.72)$$

and

$$\text{Prob}[u_T > L] = \frac{u_T(L)}{z_T - 1}. \tag{3.73}$$

In the case that $S_1(.) \neq S_2(.)$, the behavior of $U_T(z)$ is more complicated. In this case, z_B is also a (branch-point) singularity of $U_T(z)$. Two more singularities may play a role, namely the respective zeros of $z - S_1(A_T(z))$, i.e., z_T and of $z - Y_2(z)$, denoted as z_L on the real positive axis (with $z > 1$). We may thus have quite some more different cases depending on which singularities are dominant. However since it is for general $S_1(z)$ and $S_2(z)$ not clear which singularities may (or may not) be dominant, we will not go in further detail about this. For specific pgf's of the service times however the tail probabilities of the total system contents can be calculated by studying the behavior of $U_T(z)$ in the neighborhood of its dominant singularity and using Darboux's theorem (as is done in the previous chapter and as will be done in the remainder of this section).

Class-1 system contents

The dominant singularity z_H of $U_1(z)$ is the dominant zero of $z - S_1(A_1(z))$ on the positive real axis (> 1) and this singularity is a pole with multiplicity 1 (for more details see subsection 2.1.7). In the neighborhood of this pole, we approximate $U_1(z)$ by

$$U_1(z) \approx \frac{K_1}{z_H - z}, \tag{3.74}$$

with

$$K_1 = \frac{[(1-\rho_T)(A_1(z_H) - 1) + \lambda_2(S_2(A_1(z_H)) - 1)](z_H - 1)z_H}{(A_1(z_H) - 1)(S_1'(A_1(z_H))A_1'(z_H) - 1)}. \tag{3.75}$$

Using Darboux's theorem on expression (3.74), we find

$$u_1(n) = \text{Prob}[u_1 = n] \tag{3.76}$$

$$\approx K_1 z_H^{-n-1}. \tag{3.77}$$

and

$$\text{Prob}[u_1 > L] \approx \frac{K_1 z_H^{-L}}{z_H - 1}. \tag{3.78}$$

for large enough n and L.

Class-2 system contents

From expression (3.54) it is seen that $U_2(z)$ has 2 important singularities, the single pole z_L and the branch point z_B, with z_L the (dominant) zero of $z-Y_2(z)$ and z_B the (dominant) branchpoint of $Y_1(z)$ (thus with $E_1^{(1)}(Y_1(z_B), z_B) = 1$). The tail behavior of the system contents of class-2 packets is thus characterized by z_L or z_B, depending on which is the dominant (i.e., smallest) singularity. Three different types of tail behavior may thus occur, namely when z_L is dominant, when $z_L = z_B$ and when z_B is dominant. In those three cases, $U_2(z)$ is approximated in the neighborhood of its dominant singularity by:

$$U_2(z) \approx \begin{cases} \dfrac{K_2^{(1)}}{z_L - z} & \text{if } z_L \text{ dominant} \\ \dfrac{K_2^{(2)}}{(z_B - z)^{1/2}} & \text{if } z_L = z_B \text{ dominant} \\ U_2(z_B) - K_2^{(3)} (z_B - z)^{1/2} & \text{if } z_B \text{ dominant,} \end{cases} \tag{3.79}$$

with the constants $K_2^{(i)}$ equal to

$$K_2^{(1)} = \frac{(1-\rho_T) S_2(A_2(z_L))\,(z_L - 1)\,(A(Y_1(z_L), z_L) - 1)}{(Y_2'(z_L) - 1)\,(A_2(z_L) - 1)} \tag{3.80}$$

$$K_2^{(2)} = \frac{(1-\rho_T) S_2(A_2(z_B))\,(z_B - 1)\,(A(Y_1(z_B), z_B) - 1)}{K_{Y_2}\,(A_2(z_B) - 1)} \tag{3.81}$$

$$K_2^{(3)} = \frac{(1-\rho_T) S_2(A_2(z_B))\,(z_B - 1)}{(z_B - S_2(A(Y_1(z_B), z_B)))\,(A_2(z_B) - 1)} \times \left\{ K_{Y_1} A^{(1)}(Y_1(z_B), z_B) - \frac{K_{Y_2}\,(A(Y_1(z_B), z_B) - 1)}{z_B - S_2(A(Y_1(z_B), z_B))} \right\}. \tag{3.82}$$

Using Darboux's theorem once again, we find the tail probabilities of the class-2 system contents for the three possible cases:

$$u_2(n) \approx \begin{cases} K_2^{(1)} z_L^{-n-1} & \text{if } z_L \text{ dominant} \\ \dfrac{K_2^{(2)} n^{-1/2} z_B^{-n}}{\sqrt{z_B \pi}} & \text{if } z_L = z_B \text{ dominant} \\ \dfrac{K_2^{(3)}}{2} \sqrt{\dfrac{z_B}{\pi}} n^{-3/2} z_B^{-n} & \text{if } z_B \text{ dominant,} \end{cases} \tag{3.83}$$

and

$$\text{Prob}[u_2 > L] \approx \frac{u_2(L)}{z_* - 1}, \tag{3.84}$$

with z_* the dominant singularity and which is found by inverting $(U_2(z) - 1)/(z-1)$, the z-transform of the Prob$[u_2 > n]$, $n \geq 0$ (see subsection 2.1.7 for more details).

3.4 Queue contents

The *queue contents* - defined as the number of packets in the queue (thus without the possible one in the server), is easily derived from the system contents. We denote the queue contents of class-j at the beginning of the k-th slot by $q_{j,k}$ ($j = 1, 2$). We relate the queue contents at the beginning of the k-th slot to the system contents at the beginning of the last start-slot preceding this slot. We denote this start-slot by start-slot l (when slot k is a start-slot it is assumed to be start-slot l). As in subsection 3.3, the server will be in one of three states: no service during slot k, a class-1 packet is served during slot k or a class-2 packet is served during slot k. A similar analysis can thus be performed as in subsection 3.3, leading to a similar expression as expression (3.29) for the joint pgf $Q(z_1, z_2)$ of the steady-state class-1 and class-2 queue contents at the beginning of a random slot:

$$Q(z_1, z_2) = \lim_{k\to\infty} \mathrm{E}\left[z_1^{q_{1,k}} z_2^{q_{2,k}}\right] \tag{3.85}$$

$$\begin{aligned} =&U(0,0) + \frac{1-U(0,0)}{\rho_T(A(z_1,z_2)-1)} \left\{ \frac{\lambda_2(E_2(z_1,z_2)-1)}{z_2} \frac{N(0,z_2)-N(0,0)}{N(0,1)-N(0,0)} \right. \\ &\left. + \frac{\lambda_1(E_1(z_1,z_2)-1)}{z_1} \frac{N(z_1,z_2)-N(0,z_2)}{1-N(0,1)} \right\}. \end{aligned} \tag{3.86}$$

This expression varies from expression (3.29) in the second and third term: in the second (third respectively) term an extra division by z_2 (by z_1) respectively is added. This is done because the second (third respectively) term gives the partial pgf of the queue contents in case a class-2 (class-1 respectively) packet is served during slot k (and thus the queue contents of class-2 (class-1 respectively) is one less than the class-2 (class-1 respectively) system contents). Substituting $N(z_1, z_2)$ by its expression (3.13) yields

$$\begin{aligned} Q(z_1, z_2) =&(1-\rho_T)\frac{z_1-1}{z_1 - E_1(z_1,z_2)} + (1-\rho_T) \\ &\times \frac{(A(Y_1(z_2),z_2)-1)\left\{ \begin{array}{l} z_1(E_2(z_1,z_2)-1) + z_2(1-E_1(z_1,z_2)) \\ +E_1(z_1,z_2) - E_2(z_1,z_2) \end{array} \right\}}{(A(z_1,z_2)-1)(z_1-E_1(z_1,z_2))(z_2-Y_2(z_2))}. \end{aligned} \tag{3.87}$$

From this expression, marginal pgf's, moments and tail probabilities of the total, the class-1 and the class-2 queue contents can be calculated as is done in section 3.3 for the system contents.

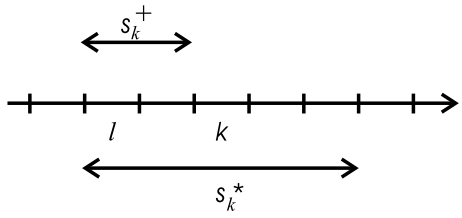


Figure 3.2: Service time of the packet in service during slot k

3.5 Unfinished work

The total unfinished work at the beginning of slot k, denoted by $w_{T,k}$, is defined as the number of slots it takes to serve all packets in the system at the beginning of slot k, when no new packets arrive from slot k on. Furthermore, the unfinished work of class-j ($j = 1, 2$) at the beginning of slot k, denoted by $w_{j,k}$, is defined as the number of slots of the total unfinished work that are effectively spent on serving class-j packets.

As opposed to single-slot service times (see section 2.3), the unfinished work and the system contents are not identical stochastic variables in this case. We define $W_k(z_1, z_2)$ as the pgf of the unfinished work at the beginning of slot k, or

$$W_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{w_{1,k}} z_2^{w_{2,k}}\right]. \tag{3.88}$$

We relate the unfinished work at the beginning of the k-th slot to the system contents at the beginning of the last start-slot preceding this slot. We denote this start-slot by start-slot l (when slot k is a start-slot it is assumed to be the l-th start-slot). As explained in subsection 3.3, the server is in one of three states: no service during slot k, a class-1 packet is served during slot k or a class-2 packet is served during slot k. We denote the service time of the packet in service during slot k by s_k^* and its elapsed part by s_k^+. The latter random variable is the amount of service that the packet served has already received at the beginning of slot k (see Figure 3.2).

The following relationships between the $w_{j,k}$ and the $n_{j,l}$ can be obtained:

1. The server is idle during slot k yielding

$$w_{j,k} = 0, \tag{3.89}$$

since the system is empty when the server is idle.

2. A class-2 packet is in service during slot k (implying that $n_{1,l} = 0, n_{2,l} > 0$):

$$w_{1,k} = \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)} \tag{3.90}$$

$$w_{2,k} = \left(s_k^* - s_k^+\right) + \sum_{m=1}^{n_{2,l}-1} \tilde{s}_{2,m} + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{2,k-i}} s_{2,m}^{(k-i)}, \tag{3.91}$$

with the $\tilde{s}_{2,m}$ the service times of the class-2 packets already in the queue at the beginning of the ongoing service (thus without the packet in service during slot k), the $s_{j,m}^{(k-i)}$'s ($1 \leq i \leq s_k^+$, $j = 1, 2$) the service times of the class-j packets that arrived during slot $k - i$. $w_{1,k}$ equals the sum of the service times of the class-1 packets that arrived during s_k^+. $w_{2,k}$ equals the sum of the residual service time of the packet in service (the first term), the superposition of the service times of the class-2 packets present in the queue at the beginning of start-slot l (the second term) and the superposition of the service times of the class-2 packets that arrived during the elapsed service time (the third term).

3. A class-1 packet is in service during slot k (i.e., $n_{1,l} > 0$), yielding

$$w_{1,k} = \left(s_k^* - s_k^+\right) + \sum_{m=1}^{n_{1,l}-1} \tilde{s}_{1,m} + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)} \tag{3.92}$$

$$w_{2,k} = \sum_{m=1}^{n_{2,l}} \tilde{s}_{2,m} + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{2,k-i}} s_{2,m}^{(k-i)}, \tag{3.93}$$

with the $\tilde{s}_{1,m}$ the service times of the class-1 packets already in the queue at the beginning of the ongoing service.

Taking the z-transform of these equations, we find

$$\begin{aligned} &W_k(z_1, z_2) \\ &\quad = \text{Prob[no service]} + \text{Prob[service class-2]} \\ &\quad \times \mathrm{E}\left[z_1^{\sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)}} z_2^{\left(s_k^* - s_k^+\right) + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{2,k-i}} s_{2,m}^{(k-i)}} | \text{service class-2} \right] \\ &\quad \times \mathrm{E}\left[z_2^{\sum_{m=1}^{n_{2,l}-1} \tilde{s}_{2,m}} | \text{service class-2} \right] \\ &\quad + \text{Prob[service class-1]} \\ &\quad \times \mathrm{E}\left[z_1^{\left(s_k^* - s_k^+\right) + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)}} z_2^{\sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{2,k-i}} s_{2,m}^{(k-i)}} | \text{service class-1} \right] \end{aligned} \tag{3.94}$$

$$\times \mathrm{E}\left[z_1^{\sum_{m=1}^{n_{1,l}-1}\tilde{s}_{1,m}} z_2^{\sum_{m=1}^{n_{2,l}}\tilde{s}_{2,m}} | \text{service class-1}\right],$$

where we specifically used the independence between the $n_{j,l}$ ($j = 1, 2$) and (s_k^*, s_k^+) (and the arrivals during s_k^+) respectively under the condition that a class-1 or class-2 service is ongoing. Since s_k^+ is the elapsed part of the service time s_k^*, both variables are mutually correlated and we define

$$S_j^*(x,z) = \lim_{k\to\infty} \mathrm{E}\left[x^{s_k^+} z^{s_k^*} | \text{service class-}j\right]. \tag{3.95}$$

We denote the steady-state version of $W_k(z_1, z_2)$ by $W(z_1, z_2)$, i.e.,

$$W(z_1,z_2) \triangleq \lim_{k\to\infty} W_k(z_1,z_2). \tag{3.96}$$

Letting $k \to \infty$, using the definitions of $S_2^*(x,z)$ and $S_1^*(x,z)$ and substituting expressions (3.24)-(3.26) and expression (3.34) in expression (3.94), yield

$$\begin{aligned} W(z_1,z_2) =& (1-\rho_T) + \rho_2 S_2^*\left(\frac{A(S_1(z_1),S_2(z_2))}{z_2}, z_2\right)\frac{N(0,S_2(z_2)) - N(0,0)}{S_2(z_2)(N(0,1)-N(0,0))} \\ &+ \rho_1 S_1^*\left(\frac{A(S_1(z_1),S_2(z_2))}{z_1}, z_1\right)\frac{N(S_1(z_1),S_2(z_2)) - N(0,S_2(z_2))}{S_1(z_1)(1-N(0,1))}, \end{aligned} \tag{3.97}$$

with $N(z_1, z_2)$ the joint pgf of the steady-state class-1 and class-2 contents at the beginning of a randomly chosen start-slot. Note that we used the fact that $n_{1,l} = 0, n_{2,l} > 0$ ($n_{1,l} > 0$ respectively) when a class-2 (class-1 respectively) service is ongoing during slot k.

We still have to calculate the $S_j^*(x,z)$ ($j = 1, 2$). First we note that

$$\text{Prob}[s_j^+ = n, s_j^* = m] = \text{Prob}[s_j^+ = n | s_j^* = m]\text{Prob}[s_j^* = m]. \tag{3.98}$$

Taking into account that an arbitrary class-j service slot is more likely to be chosen in a larger service time, we find

$$\text{Prob}[s_j^* = m] = \frac{m s_j(m)}{\mu_j}, \tag{3.99}$$

with $s_j(m)$ the pmf of the service times of class-j packets. Furthermore it is easily seen that

$$\text{Prob}[s_j^+ = n | s_j^* = m] = \frac{1}{m}, \tag{3.100}$$

for $n = 0, \ldots, m-1$. Expression (3.98) thus becomes

$$\text{Prob}[s_j^+ = n, s_j^* = m] = \frac{s_j(m)}{\mu_j}, \tag{3.101}$$

for $n = 0, \ldots, m-1$ (and 0 for $n \geq m$). Taking the (two-dimensional) z-transform of this expression leads to

$$S_j^*(x,z) = \frac{S_j(xz) - S_j(z)}{\mu_j(x-1)}, \tag{3.102}$$

for $j = 1, 2$. Note that $S_j^+(z)$ defined in subsection 3.3.1 equals the marginal pgf $S_j^*(z,1)$. Thus it can be seen from (3.102) that expression (3.28) is indeed valid.

Substituting this expression and expression (3.13) in (3.97), we find a fully determined version for $W(z_1, z_2)$:

$$\begin{aligned} W(z_1,z_2) =& (1-\rho_T)\frac{A(S_1(z_1),S_2(z_2))(z_1-1)}{z_1 - A(S_1(z_1),S_2(z_2))} + (1-\rho_T) \\ & \times \frac{A(S_1(z_1),S_2(z_2))(z_1-z_2)(A(Y_1(S_2(z_2)),S_2(z_2))-1)}{(z_1 - A(S_1(z_1),S_2(z_2)))(z_2 - A(S_1(z_1),S_2(z_2)))} \\ & \times \frac{S_2(z_2) - E_2(S_1(z_1),S_2(z_2))}{S_2(z_2) - Y_2(S_2(z_2))}. \end{aligned} \tag{3.103}$$

From this joint pgf, marginal pgf's and performance measures can be calculated as is done in section 2.1. It is seen that substituting $S_1(z) = S_2(z) = z$ in this equation yields the corresponding expression of $W(z_1, z_2)$ in chapter 2 (expression (2.86)).

3.6 Packet delay

The packet delay is defined as the total time period a tagged packet spends in the system, or more precisely, the number of slots between the end of the packet's arrival slot and the end of its departure slot. As before, we denote the steady-state packet delay of a tagged class-j packet by d_j and its pgf by $D_j(z)$ ($j = 1, 2$). Furthermore, the steady-state packet delay of a *random* packet is denoted by d with pgf $D(z)$. Before deriving expressions for $D_1(z)$, $D_2(z)$ and $D(z)$, we first define some notions and stochastic variables we will frequently use in this section.

In the next subsections, we tag a packet. We denote the arrival slot of the tagged packet by slot k. If slot k is a start-slot, it is assumed to be start-slot l. If slot k is not a start-slot on the other hand, the last start-slot preceding slot k is

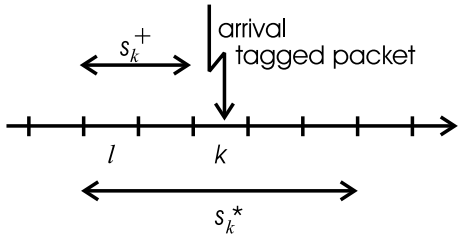


Figure 3.3: Service time of the packet in service during the tagged packet's arrival slot

assumed to be start-slot l. We denote the number of class-j packets that arrive during slot k, but which are served before the tagged packet by $f_{j,k}$ ($j = 1, 2$). We furthermore denote the service time of the tagged class-j packet by $\hat{s}_j$ ($j = 1, 2$). We finally denote the service time and the elapsed service time of the packet in service (if any) during the arrival slot of the tagged packet by s_k^* and s_k^+ respectively (see Figure 3.3). Note that these latter two stochastic variables have the same distributions as the stochastic variables with the same notation in section 3.5 (see Figure 3.2), since the arrival slot of a random tagged packet is basically identical to a random slot (for i.i.d. per-slot arrivals). Therefore, we use the same notations in this section as in the previous section.

3.6.1 Pgf $D_1(z)$ of the class-1 packet delay

We tag a class-1 packet. There are 3 possibilities when the tagged packet arrives:

1. The server is idle during slot k, yielding

$$d_1 = \sum_{m=1}^{f_{1,k}} s_{1,m}^{(k)} + \hat{s}_1, \tag{3.104}$$

   with the $s_{1,m}^{(k)}$'s the service times of the class-1 packets that arrived during slot k, but that are served before the tagged class-1 packet.

2. A class-2 packet is in service during slot k, yielding

$$d_1 = (s_k^* - s_k^+ - 1) + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)} + \sum_{m=1}^{f_{1,k}} s_{1,m}^{(k)} + \hat{s}_1, \tag{3.105}$$

   with $s_{1,m}^{(k-i)}$ ($0 \leq i \leq s_k^+$) the service time of the m-th class-1 packet that arrived during slot $k-i$. The residual service time of the packet in service

during slot k contributes in the first term, the service times of the class-1 packets in the system at the beginning of slot k contribute in the second term, the service times of the class-1 packets arrived during slot k, but served before the tagged class-1 packet contribute in the third term, and finally the service time of the tagged class-1 packet itself contributes in the last term.

3. A class-1 packet is in service during slot k yielding

$$d_1 = (s_k^* - s_k^+ - 1) + \sum_{m=1}^{n_{1,l}-1} \tilde{s}_{1,m} + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)} + \sum_{m=1}^{f_{1,k}} s_{1,m}^{(k)} + \hat{s}_1. \tag{3.106}$$

The difference with the previous situation is that there may be multiple class-1 packets in the buffer (apart from the one in service) at the beginning of start-slot l, which contribute to the tagged packet's delay. If we denote by $\tilde{s}_{1,m}$ the service times of the class-1 packets already in the queue at the beginning of the ongoing service (thus without the packet in service during slot k), then this condition is quantified by the second term in the right-hand side of the above expression.

Using these equations, we derive an expression for $D_1(z)$:

$$\begin{aligned} D_1(z) =& \mathrm{E}\left[z^{d_1}\{\text{no service}\}\right] + \mathrm{E}\left[z^{d_1}\{\text{service class-2}\}\right] \qquad (3.107)\\ &+ \mathrm{E}\left[z^{d_1}\{\text{service class-1}\}\right] \\ =& \mathrm{E}\left[z^{\sum_{m=1}^{f_{1,k}} s_{1,m}^{(k)} + \hat{s}_1}\right] \Bigg\{ 1 - \rho_T \qquad (3.108)\\ &+ \rho_2 \mathrm{E}\left[z^{s_k^* - s_k^+ - 1 + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)}} | \text{service class-2}\right] \\ &+ \rho_1 \mathrm{E}\left[z^{s_k^* - s_k^+ - 1 + \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{1,k-i}} s_{1,m}^{(k-i)}} | \text{service class-1}\right] \\ &\times \mathrm{E}\left[z^{\sum_{m=1}^{n_{1,l}-1} \tilde{s}_{1,m}} | \text{service class-1}\right] \Bigg\}, \end{aligned}$$

where we used that Prob[no service]=$1 - \rho_T$, Prob[service class-2]=ρ_2 and Prob[service class-1]=ρ_1 (see section 3.3). In the third term, we also used the independence of the number of class-1 packets in the queue at the beginning of start-slot l and the number of packets arriving between start-slots l and $l+1$ (under the condition that a class-1 packet is being served between the start-slots). This expression can be further transformed by keeping in mind that $n_{1,l} = 0, n_{2,l} > 0$ ($n_{1,l} > 0$ respectively) when a service of a class-2 (class-1

respectively) packet is ongoing, yielding

$$D_1(z) = F_1^{(1)}(S_1(z))S_1(z)\left\{1-\rho_T+\rho_2\frac{S_2^*(\frac{A_1(S_1(z))}{z},z)}{z}\right. \\ \left.+\rho_1\frac{S_1^*(\frac{A_1(S_1(z))}{z},z)}{z}\frac{N(S_1(z),1)-N(0,1)}{S_1(z)(1-N(0,1))}\right\}, \tag{3.109}$$

with

$$F_1^{(1)}(z) \triangleq \lim_{k\to\infty} \mathrm{E}[z^{f_{1,k}}] \tag{3.110}$$

$$S_2^*(x,z) \triangleq \lim_{k\to\infty} \mathrm{E}[x^{s_k^+} z^{s_k^*} | \text{service class-2}] \tag{3.111}$$

$$S_1^*(x,z) \triangleq \lim_{k\to\infty} \mathrm{E}[x^{s_k^+} z^{s_k^*} | \text{service class-1}]. \tag{3.112}$$

$F_1^{(1)}(z)$ was already calculated in subsection 2.4.1:

$$F_1^{(1)}(z) = \frac{A_1(z)-1}{\lambda_1(z-1)}, \tag{3.113}$$

and the $S_j^*(x,z)$ $(j=1,2)$ were already calculated in section 3.5:

$$S_j^*(x,z) = \frac{S_j(xz)-S_j(z)}{\mu_j(x-1)}. \tag{3.114}$$

We now obtain a fully determined expression for $D_1(z)$ from equation (3.109) together with equations (3.13), (3.113) and (3.114):

$$D_1(z) = \frac{1-\rho_1}{\lambda_1}\frac{S_1(z)(z-1)}{z-A_1(S_1(z))}\frac{A_1(S_1(z))-1}{S_1(z)-1}\left(\frac{1-\rho_T}{1-\rho_1}+\frac{\rho_2}{1-\rho_1}\frac{S_2(z)-1}{\mu_2(z-1)}\right). \tag{3.115}$$

Note that the *stochastic decomposition property* discussed in subsection 3.3.3 for the class-1 system contents, is also apparent in this expression of the pgf of the delay of the class-1 packets: the first factor of expression (3.115) equals the pgf of the packet delay in a *single-class* system with a FIFO scheduling discipline and the second factor equals the influence of the class-2 packets on the pgf of the class-1 packet delay. This influence is zero with probability $(1-\rho_T)/(1-\rho_1)$ and equals a residual class-2 service time with probability $\rho_2/(1-\rho_1)$.

3.6.2 Pgf $D_2(z)$ of the class-2 packet delay

Because of the priority discipline, finding an expression for $D_2(z)$ will be a bit more involved. We now tag a class-2 packet that enters the buffer during slot k. We will again use the notion of *sub-busy periods* to analyze the class-2 packet delay, as in the previous chapter. In this case, we define two kinds of sub-busy periods, i.e., *sub-busy periods initiated by a class-1 packet* and *sub-busy periods initiated by a class-2 packet*.

The first type is defined as follows: it starts at the beginning of the slot the initiating class-1 packet enters the server. Assume that at that time instant m other class-1 packets are present in the system. The sub-busy period then ends at the beginning of the slot where - for the first time - the number of class-1 packets in the system equals $m-1$, i.e., one less than at the beginning of the sub-busy period. Or in other words, the sub-busy period initiated by a class-1 packet is the period necessary to decrease the number of class-1 packets in the system by 1.

The second type - a sub-busy period initiated by a class-2 packet - is defined as follows: it starts at the beginning of the slot the initiating class-2 packet enters the server and it ends at the beginning of which a new class-2 packet can enter the server (if there are any). It thus ends when the class-2 packet left the system *and* when the system is emptied of class-1 packets (for the first time).

In order to analyze the delay of a tagged class-2 packet, let us refer to the packets in the system at the end of slot k, but that have to be served before the tagged packet as the "primary packets". So, basically, the tagged class-2 packet can enter the server, when *all primary packets* and *all class-1 packets that arrived after slot k* (i.e., while the tagged packet is waiting in the queue) are transmitted. To summarize, all primary class-j packets will add a class-j sub-busy period to the delay of the tagged packet. Let $\tilde{v}_{j,m}$ denote the length of the m-th class-j sub-busy period added to the tagged packet's delay by the m-th class-j packet already in the queue at the beginning of start-slot l and let $v_{j,m}^{(i)}$ denote the length of the sub-busy period added to the delay of the tagged class-2 packet by the m-th class-j packet that arrived during slot i.

When the tagged class-2 packet arrives, the server can - again - be in one of three possible states: the server is idle, a class-2 packet is in service or a class-1 packet is in service. In the following, we give the expressions for d_2 in the three situations.

1. The server is idle during slot k, yielding

$$d_2 = \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \hat{s}_2, \tag{3.116}$$

with $f_{j,k}$ the number of class-j packets arriving in the same slot as the tagged packet but that are served before it. $f_{1,k}$ class-1 primary packets and $f_{2,k}$ class-2 primary packets that arrived during slot k have to be served before the tagged class-2 packet and those primary packets all add a sub-busy period to the tagged packet's delay.

2. A class-2 packet is in service during slot k, yielding

$$d_2 = (s_k^* - s_k^+ - 1) + \sum_{j=1}^{2} \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{j,k-i}} v_{j,m}^{(k-i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \sum_{i=1}^{s_k^*-s_k^+-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{m=1}^{n_{2,l}-1} \tilde{v}_{2,m} + \hat{s}_2. \tag{3.117}$$

The residual service time of the packet in service during slot k contributes in the first term; the sub-busy periods added to the tagged packet's delay by the class-1 and class-2 packets that arrived during the elapsed service time contribute in the second term; the sub-busy periods added to the tagged packet's delay by the class-1 and class-2 packets arriving during slot k, but that have to be served before the tagged class-2 packet contribute in the third term; the sub-busy periods added to the tagged packet's delay by the class-1 packets arriving during the residual service time contribute in the fourth term; the sub-busy periods added to the tagged packet's delay by the class-2 packets already in the queue at the beginning of start-slot l contribute in the fifth term; finally the service time of the tagged class-2 packet itself contributes in the last term.

3. A class-1 packet is in service during slot k, yielding

$$d_2 = (s_k^* - s_k^+ - 1) + \sum_{m=1}^{n_{1,l}-1} \tilde{v}_{1,m} + \sum_{j=1}^{2} \sum_{i=1}^{s_k^+} \sum_{m=1}^{a_{j,k-i}} v_{j,m}^{(k-i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \sum_{i=1}^{s_k^*-s_k^+-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{m=1}^{n_{2,l}} \tilde{v}_{2,m} + \hat{s}_2. \tag{3.118}$$

The difference with the previous situation is that there may be multiple class-1 packets in the buffer (apart from the one in service) at the beginning of start-slot l, which will contribute to the tagged packet's delay. This condition is quantified by the second term in the right-hand side of the above expression.

Using equations (3.116)-(3.118), we derive an expression for $D_2(z)$:

$$
\begin{aligned}
D_2(z) =& \mathrm{E}\left[z^{d_2}\{\text{no service}\}\right] + \mathrm{E}\left[z^{d_2}\{\text{service class-2}\}\right] && (3.119)\\
&+ \mathrm{E}\left[z^{d_2}\{\text{service class-1}\}\right]\\
=& \mathrm{E}\left[z^{\sum_{j=1}^{2}\sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)}+\hat{s}_2}\right]\Bigg\{1-\rho_T && (3.120)\\
&+\rho_2 \mathrm{E}\left[z^{\begin{pmatrix} s_k^*-s_k^++1+\sum_{j=1}^{2}\sum_{i=1}^{s_k^+}\sum_{m=1}^{a_{j,k-i}} v_{j,m}^{(k-i)} \\ +\sum_{i=1}^{s_k^*-s_k^+-1}\sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} \end{pmatrix}}\middle|\text{service class-2}\right]\\
&\times \mathrm{E}\left[z^{\sum_{m=1}^{n_{2,l}-1}\tilde{v}_{2,m}}|\text{service class-2}\right]\\
&+\rho_1 \mathrm{E}\left[z^{\begin{pmatrix} s_k^*-s_k^++1+\sum_{j=1}^{2}\sum_{i=1}^{s_k^+}\sum_{m=1}^{a_{j,k-i}} v_{j,m}^{(k-i)} \\ +\sum_{i=1}^{s_k^*-s_k^+-1}\sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} \end{pmatrix}}\middle|\text{service class-1}\right]\\
&\times \mathrm{E}\left[z^{\sum_{m=1}^{n_{1,l}-1}\tilde{v}_{1,m}+\sum_{m=1}^{n_{2,l}}\tilde{v}_{2,m}}|\text{service class-1}\right]\Bigg\}.
\end{aligned}
$$

It is clear that the lengths of the sub-busy periods initiated by class-1 packets are i.i.d. and thus all have the same pgf $V_1(z)$. Also the lengths of the sub-busy periods initiated by class-2 packets are i.i.d., and their common pgf is denoted by $V_2(z)$. Notice that $f_{1,k}$ and $f_{2,k}$ are correlated, since they depend on $a_{1,k}$ and $a_{2,k}$ respectively, which are assumed to be correlated throughout this dissertation. Equation (3.120) then becomes

$$
\begin{aligned}
D_2(z) =& \mathrm{E}\left[V_1(z)^{f_{1,k}}V_2(z)^{f_{2,k}}\right] S_2(z)\Bigg\{1-\rho_T && (3.121)\\
&+\rho_2 \frac{\mathrm{E}\left[\left(\dfrac{A(V_1(z),V_2(z))}{zA_1(V_1(z))}\right)^{s_k^+}(zA_1(V_1(z)))^{s_k^*}|\text{service class-2}\right]}{zA_1(V_1(z))}\\
&\times \frac{\mathrm{E}\left[V_2(z)^{n_{2,l}}|n_{1,l}=0,n_{2,l}>0\right]}{V_2(z)}\\
&+\rho_1 \frac{\mathrm{E}\left[\left(\dfrac{A(V_1(z),V_2(z))}{zA_1(V_1(z))}\right)^{s_k^+}(zA_1(V_1(z)))^{s_k^*}|\text{service class-1}\right]}{zA_1(V_1(z))}\\
&\frac{\mathrm{E}\left[V_1(z)^{n_{1,l}}V_2(z)^{n_{2,l}}|n_{1,l}>0\right]}{V_1(z)}\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
=&F^{(2)}(V_1(z),V_2(z))S_2(z)\Bigg\{1-\rho_T \qquad (3.122)\\
&+\rho_2\frac{S_2^*\left(\frac{A(V_1(z),V_2(z))}{zA_1(V_1(z))},zA_1(V_1(z))\right)}{zA_1(V_1(z))}\frac{N(0,V_2(z))-N(0,0)}{(N(0,1)-N(0,0))V_2(z)}\\
&+\rho_1\frac{S_1^*\left(\frac{A(V_1(z),V_2(z))}{zA_1(V_1(z))},zA_1(V_1(z))\right)}{zA_1(V_1(z))}\\
&\times\frac{N(V_1(z),V_2(z))-N(0,V_2(z))}{(1-N(0,1))V_1(z)}\Bigg\},
\end{aligned}
$$

with

$$F^{(2)}(z_1,z_2)\triangleq\mathrm{E}\left[z_1^{f_{1,k}}z_2^{f_{2,k}}\right], \qquad (3.123)$$

and

$$S_j^*(x,z)\triangleq\lim_{k\to\infty}\mathrm{E}\left[x^{s_k^+}z^{s_k^*}|\text{service class-}j\right], \qquad (3.124)$$

with $j=1,2$. As in the previous chapter, the random variables $f_{1,k}$ and $f_{2,k}$ can be shown to have the following joint pgf:

$$F^{(2)}(z_1,z_2)=\frac{A(z_1,z_2)-A_1(z_1)}{\lambda_2(z_2-1)}. \qquad (3.125)$$

The $S_j^*(x,z)$'s ($j=1,2$) are already calculated in the previous subsection and are given by equation (3.102).

Finally, we find expressions for $V_1(z)$ and $V_2(z)$. Denote a (random) sub-busy period of class-j by v_j and the service time of the class-j packet initiating the sub-busy period by s_j, $j=1,2$. Furthermore, denoting the class-1 arrivals during the m-th slot of this service time by $a_1^{(m)}$, we have

$$v_j=s_j+\sum_{m=1}^{s_j}\sum_{l=1}^{a_1^{(m)}}v_{1,l}^{(m)}, \qquad (3.126)$$

with $v_{1,l}^{(m)}$ defined as the class-1 sub-busy period added to the original sub-busy period by the l-th class-1 packet arriving in the m-th slot of the service time s_j. Since v_j and s_j have pgf's $V_j(z)$ and $S_j(z)$ respectively and since the class-1 sub-busy periods $v_{1,l}^{(m)}$ are i.i.d. and have a common pgf $V_1(z)$, z-

transforming (3.126) yields

$$V_j(z) = S_j(zA_1(V_1(z))), \tag{3.127}$$

with $j = 1, 2$.

Equation (3.122) together with equations (3.13), (3.102) and (3.125) leads to:

$$D_2(z) = \frac{1-\rho_T}{\lambda_2} \frac{S_2(z)(A(V_1(z), V_2(z)) - A_1(V_1(z)))}{zA_1(V_1(z)) - A(V_1(z), V_2(z))} \frac{zA_1(V_1(z)) - 1}{V_2(z) - 1}, \tag{3.128}$$

with $V_j(z)$ ($j = 1, 2$) implicitly given by equation (3.127).

In the special case of deterministic service times of 1 slot for both classes, expression (3.128) equals expression (2.104).

3.6.3 Pgf $D(z)$ of the delay of a random packet

In this subsection, we derive the pgf $D(z)$ of a *random* packet arriving in the system. Tagging a random arriving packet, it is of class-1 with probability λ_1/λ_T and of class-2 with probability λ_2/λ_T. We thus get

$$D(z) = \frac{\lambda_1}{\lambda_T} D_1(z) + \frac{\lambda_2}{\lambda_T} D_2(z). \tag{3.129}$$

Substituting expression (3.115) and (3.128) in this expression leads to

$$\begin{aligned} D(z) = &\frac{1-\rho_1}{\lambda_T} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1} \left(\frac{1-\rho_T}{1-\rho_1} + \frac{\rho_2}{1-\rho_1} \frac{S_2(z) - 1}{\mu_2(z-1)} \right) \\ &+ \frac{1-\rho_T}{\lambda_T} \frac{S_2(z)(A(V_1(z), V_2(z)) - A_1(V_1(z)))}{zA_1(V_1(z)) - A(V_1(z), V_2(z))} \frac{zA_1(V_1(z)) - 1}{V_2(z) - 1}. \end{aligned} \tag{3.130}$$

3.6.4 The functions $Y_j(z)$ revisited

The functions $Y_j(z)$ ($j = 1, 2$) are the pgf's of stochastic variables, namely the stochastic variables y_j, which are defined as the number of class-2 packets that arrive during a sub-busy period initiated by a class-j packet. If at the beginning of a random slot a class-j packet with service time s_j enters the server, a new sub-busy period starts (with length denoted by v_j). Denoting the number of class-2 packets that arrive during this sub-busy period by y_j

and denoting the number of class-i arrivals during the m-th slot of this sub-busy period by $a_i^{(m)}$ ($i = 1, 2$), we get

$$y_j = \sum_{m=1}^{v_j} a_2^{(m)} \tag{3.131}$$

$$= \sum_{m=1}^{s_j} \left(a_2^{(m)} + \sum_{l=1}^{a_1^{(m)}} y_{1,l}^{(m)} \right), \tag{3.132}$$

with $y_{1,l}^{(m)}$ the number of class-2 packets that arrive during the sub-busy period added to v_j by the l-th class-1 packet that arrives during the m-th slot of the service time s_j. Naturally, all $y_{1,l}^{(m)}$ have the same distribution as y_1 (since the lengths of class-1 sub-busy periods are all identically distributed) and therefore the pgf of y_j is thus indeed given by

$$Y_j(z) = S_j(A(Y_1(z), z)), \tag{3.133}$$

for $j = 1, 2$, as immediately follows from (3.132) assuming that a stationary regime is reached.

When the number of class-1 and class-2 arrivals are independent stochastic variables, it can be seen that v_j and the $a_2^{(m)}$ are also independent variables. From expression (3.131), it then follows that

$$Y_j(z) = V_j(A_2(z)), \tag{3.134}$$

for $j = 1, 2$. Note that this latter expression is not generally valid when the number of class-1 and class-2 arrivals in a slot are correlated, since v_j and the $a_2^{(m)}$'s both depend on the $a_1^{(m)}$'s in this case.

3.6.5 Calculation of moments

Taking the first derivative of equation (3.115) and substituting z by 1 yields the mean class-1 packet delay:

$$\mathrm{E}[d_1] = \frac{\mu_1}{2} + \frac{\mu_1 \mathrm{Var}[a_1]}{2\lambda_1(1-\rho_1)} + \frac{\lambda_1 \mathrm{Var}[s_1]}{2(1-\rho_1)} + \frac{\lambda_2(\mu_2(\mu_2-1) + \mathrm{Var}[s_2])}{2(1-\rho_1)}. \tag{3.135}$$

Note that this expression is always at least equal to μ_1. Since $\mathrm{Var}[a_1] \geq \lambda_1(1-\lambda_1)$ (for more details see subsection 2.1.6), the second term of (3.135) equals $\mu_1/2$ for $\lambda_1 \to 0$ and $\lambda_2 \to 0$ - and keeping $\mathrm{Var}[a_1] = \lambda_1(1-\lambda_1)$ -, thus $\mathrm{E}[d_1] \geq \mu_1$. This is also intuitively clear, since the service time of a packet is an integral part of the delay of that packet. Thus the mean delay is at least the

mean service time of a packet, i.e., μ_1 in this case. It can furthermore be seen from expressions (3.59) and (3.135) that $\mathrm{E}[u_1] = \lambda_1 \mathrm{E}[d_1]$, which is also found from (the discretized version of) Little's law.

The mean delay of a class-2 packet is found by taking the first derivative of expression (3.128) and substituting z by 1, yielding

$$\begin{aligned}\mathrm{E}[d_2] =& \frac{\mu_2}{2} + \frac{\mu_1^2 \mathrm{Var}[a_1]}{2(1-\rho_T)(1-\rho_1)} + \frac{\mu_2 \mathrm{Var}[a_2]}{2\lambda_2(1-\rho_T)} + \frac{\mu_1 \mathrm{Cov}[a1, a2]}{\lambda_2(1-\rho_T)} \\ &+ \frac{\lambda_1 \mathrm{Var}[s_1] + \lambda_2 \mathrm{Var}[s_2]}{2(1-\rho_T)(1-\rho_1)} - \frac{\rho_1(\mu_2 - 1)}{2(1-\rho_1)}.\end{aligned} \tag{3.136}$$

This expression is at least μ_2, since for given mean arrival rates of class-1 and class-2 (λ_1 and λ_2 respectively), the minimal value of $\mathrm{Var}[a_j]$ is $\lambda_j(1-\lambda_j)$ ($j = 1, 2$) and the minimal value of $\mathrm{Cov}[a_1, a_2]$ is $-\lambda_1\lambda_2$. These least possible values occur when one class-1 packet and no class-2 packets arrive during a slot with probability λ_1, no class-1 packets and one class-2 packet arrive with probability λ_2 and no arrivals occur during a slot with probability $1-\lambda_T$. With these values, expression (3.136) equals μ_2 for λ_1 and λ_2 going to 0. Little's law is again valid: $\mathrm{E}[u_2] = \lambda_2 \mathrm{E}[d_2]$.

Finally, the mean delay of a random packet is given by

$$\mathrm{E}[d] = \frac{\lambda_1}{\lambda_T}\mathrm{E}[d_1] + \frac{\lambda_2}{\lambda_T}\mathrm{E}[d_2] \tag{3.137}$$

$$\begin{aligned}=& \frac{\rho_T}{2\lambda_T} + \frac{\mu_1 \mathrm{Var}[a_T]}{2\lambda_T(1-\rho_T)} - \frac{\mu_1\lambda_2(\mu_2-\mu_1)\mathrm{Var}[a_1]}{2\lambda_T(1-\rho_T)(1-\rho_1)} + \frac{(\mu_2-\mu_1)\mathrm{Var}[a_2]}{2\lambda_T(1-\rho_T)} \\ &+ \frac{(\lambda_1 \mathrm{Var}[s_1] + \lambda_2 \mathrm{Var}[s_2])(\lambda_1(1-\rho_1) + \lambda_2(1-\mu_2\lambda_1))}{2\lambda_T(1-\rho_T)(1-\rho_1)} \\ &+ \frac{\lambda_1\lambda_2(\mu_2-1)(\mu_2-\mu_1)}{2\lambda_T(1-\rho_1)}.\end{aligned} \tag{3.138}$$

We have used expressions (3.135) and (3.136) to obtain this formula. Expressions (3.56) and (3.138) clearly satisfy Little's law.

3.6.6 Calculation of tail probabilities

In this subsection, we calculate the tail probabilities of the delay of a class-1 packet, of the delay of a class-2 packet and of the delay of a random packet.

Class-1 packet delay

The dominant singularity of $D_1(z)$ (expression (3.115)), denoted by $\hat{z}_H$, is the dominant zero of $z - A_1(S_1(z))$ on the positive real axis (> 1). It is a single pole

and we can thus approximate the tail behavior of the delay of class-1 packets by

$$d_1(n) \triangleq \text{Prob}[d_1 = n] \tag{3.139}$$

$$\approx \hat{K}_1 z_H^{-n-1}, \tag{3.140}$$

and

$$\text{Prob}[d_1 > D] \approx \frac{\hat{K}_1 \hat{z}_H^{-D}}{\hat{z}_H - 1}. \tag{3.141}$$

for large enough n and D respectively, with

$$\hat{K}_1 = \frac{S_1(\hat{z}_H)\,(\hat{z}_H - 1)\,[(1-\rho_T)\,(\hat{z}_H - 1) + \lambda_2\,(S_2(\hat{z}_H) - 1)]}{\lambda_1\,(S_1(\hat{z}_H) - 1)\,(A_1'(S_1(\hat{z}_H))S_1'(\hat{z}_H) - 1)}. \tag{3.142}$$

Behavior of $V_1(z)$ and $V_2(z)$ in the neighborhood of their dominant branch-point

The tail behavior of the delay of class-2 packets is again a bit more involved because of the appearance of the function $V_1(z)$ in (3.128), which is (generally) only implicitly known. We will first observe the behavior of $V_1(z)$ and $V_2(z)$. The first derivative of $V_1(z)$ is given by

$$V_1'(z) = \frac{S_1'(zA_1(V_1(z)))A_1(V_1(z))}{1 - zS_1'(zA_1(V_1(z)))A_1'(V_1(z))}, \tag{3.143}$$

which, similar as before, indicates that $V_1(z)$ has a branch point $\hat{z}_B$, with

$$\hat{z}_B S_1'(\hat{z}_B A_1(V_1(\hat{z}_B)))A_1'(V_1(\hat{z}_B)) = 1. \tag{3.144}$$

In the neighborhood of $\hat{z}_B$, $V_1(z)$ is approximately given by

$$V_1(z) \approx V_1(\hat{z}_B) - K_{V_1}\sqrt{\hat{z}_B - z}, \tag{3.145}$$

with

$$K_{V_1} = \sqrt{\frac{2A_1(V_1(\hat{z}_B))}{\hat{z}_B A_1''(V_1(\hat{z}_B)) + S_1''(\hat{z}_B A_1(V_1(\hat{z}_B)))(\hat{z}_B A_1'(V_1(\hat{z}_B)))^3}}. \tag{3.146}$$

$\hat{z}_B$ is also a branch-point of $V_2(z)$, since

$$V_2'(z) = S_2'(zA_1(V_1(z)))\,[A_1(V_1(z)) + zA_1'(V_1(z))V_1'(z)]\,. \tag{3.147}$$

Thus in the neighborhood of $\hat{z}_B$, we find

$$V_2(z) \approx V_2(\hat{z}_B) - K_{V_2}\sqrt{\hat{z}_B - z}, \tag{3.148}$$

with

$$K_{V_2} = K_{V_1} S_2'(\hat{z}_B A_1(V_1(\hat{z}_B)))\hat{z}_B A_1'(V_1(\hat{z}_B)). \tag{3.149}$$

Class-2 packet delay

$\hat{z}_B$ is also a branch-point singularity of $D_2(z)$. A second (potential) singularity of $D_2(z)$ is given by the dominant zero $\hat{z}_L$ of $zA_1(V_1(z)) - A(V_1(z), V_2(z))$. So, $D_2(z)$ is approximated in the neighborhood of its dominant singularity by:

$$D_2(z) \approx \begin{cases} \dfrac{\hat{K}_2^{(1)}}{\hat{z}_L - z} & \text{if } \hat{z}_L \text{ dominant} \\ \dfrac{\hat{K}_2^{(2)}}{\sqrt{\hat{z}_B - z}} & \text{if } \hat{z}_L = \hat{z}_B \text{ dominant} \\ D_2(\hat{z}_B) - \hat{K}_2^{(3)}\sqrt{\hat{z}_B - z} & \text{if } \hat{z}_B \text{ dominant,} \end{cases} \tag{3.150}$$

with the constants $\hat{K}_2^{(i)}$ $(i = 1, 2, 3)$ found by investigating $D_2(z)$ in the neighborhood of its dominant singularity, yielding

$$\hat{K}_2^{(1)} = \frac{(1-\rho_T)S_2(\hat{z}_L)A_1(V_1(\hat{z}_L))\,(\hat{z}_L - 1)\,(\hat{z}_L A_1(V_1(\hat{z}_L)) - 1)}{\lambda_2\,(V_2(\hat{z}_L) - 1)\left(\dfrac{dA(V_1(z), V_2(z))}{dz} - \dfrac{d(zA_1(V_1(z))}{dz}\right)\Bigg|_{z=\hat{z}_L}} \tag{3.151}$$

$$\begin{aligned} \hat{K}_2^{(2)} &= \frac{(1-\rho_T)\,(\hat{z}_B A_1(V_1(\hat{z}_B)) - 1)}{\lambda_2(V_2(\hat{z}_B) - 1)} \\ &\times \frac{S_2(\hat{z}_B)A_1(V_1(\hat{z}_B))\,(\hat{z}_B - 1)}{K_{V_1}\left(A^{(1)}(V_1(\hat{z}_B), V_2(\hat{z}_B)) - \hat{z}_B A_1'(V_1(\hat{z}_B))\right) + K_{V_2}A^{(2)}(V_1(\hat{z}_B), V_2(\hat{z}_B))} \end{aligned} \tag{3.152}$$

$$\begin{aligned} \hat{K}_2^{(3)} &= \frac{(1-\rho_T)S_2(\hat{z}_B)}{\lambda_2(\hat{z}_B A_1(V_1(\hat{z}_B)) - A(V_1(\hat{z}_B), V_2(\hat{z}_B)))^2(V_2(\hat{z}_B) - 1)^2} \\ &\times \Big\{\Big[A_1(V_1(\hat{z}_B))\left(K_{V_1}A^{(1)}(V_1(\hat{z}_B), V_2(\hat{z}_B)) + K_{V_2}A^{(2)}(V_1(\hat{z}_B), V_2(\hat{z}_B))\right) \\ &- K_{V_1}A(V_1(\hat{z}_B), V_2(\hat{z}_B))A_1'(V_1(\hat{z}_B))\Big](\hat{z}_B - 1)(\hat{z}_B A_1(V_1(\hat{z}_B)) - 1) \\ &\times (V_2(\hat{z}_B) - 1) + [K_{V_1}\hat{z}_B A_1'(V_1(\hat{z}_B))(V_2(\hat{z}_B) - 1) - K_{V_2}(\hat{z}_B A_1(V_1(\hat{z}_B)) - 1)] \\ &\times (A(V_1(\hat{z}_B), V_2(\hat{z}_B)) - A_1(V_1(\hat{z}_B)))\,(\hat{z}_B A_1(V_1(\hat{z}_B)) - A(V_1(\hat{z}_B), V_2(\hat{z}_B)))\Big\}. \end{aligned} \tag{3.153}$$

Using Darboux's theorem, we find

$$d_2(n) \triangleq \text{Prob}[d_2 = n] \tag{3.154}$$

$$\approx \begin{cases} \hat{K}_2^{(1)} \hat{z}_L^{-n-1} & \text{if } \hat{z}_L \text{ dominant} \\ \dfrac{\hat{K}_2^{(2)} n^{-1/2} \hat{z}_B^{-n}}{\sqrt{\hat{z}_B \pi}} & \text{if } \hat{z}_L = \hat{z}_B \text{ dominant} \\ \dfrac{\hat{K}_2^{(3)}}{2} \sqrt{\dfrac{\hat{z}_B}{\pi}} n^{-3/2} \hat{z}_B^{-n} & \text{if } \hat{z}_B \text{ dominant,} \end{cases} \tag{3.155}$$

and

$$\text{Prob}[d_2 > D] \approx \frac{d_2(D)}{\hat{z}_* - 1}, \tag{3.156}$$

with $\hat{z}_*$ the dominant singularity of $D_2(z)$.

Delay random packet

Finally, we calculate the tail probabilities of the delay of a random packet. Its pgf $D(z)$ is given by expression (3.130). The dominant singularity of this function is $\hat{z}_L$ or $\hat{z}_B$, depending on which is smallest.

Note that - similar as in the previous chapter - the dominant singularity of $D_1(z)$ - $\hat{z}_H$ - is also a singularity of $D(z)$, but $\hat{z}_H$ is never dominant, since $\hat{z}_H > \hat{z}_B$. We will first prove this latter inequality. Firstly, since $\hat{z}_B$ fulfils the relation $S_1'(\hat{z}_B A_1(V_1(\hat{z}_B)))\hat{z}_B A_1'(V_1(\hat{z}_B)) = 1$ and since $\hat{z}_B$ is larger than 1

$$S_1'(\hat{z}_B A_1(V_1(\hat{z}_B)))A_1'(V_1(\hat{z}_B)) < 1, \tag{3.157}$$

or - by using the definition of $V_1(z) = S_1(zA_1(V_1(z)))$ -

$$\left.\frac{dA_1(S_1(z))}{dz}\right|_{z=\hat{z}_B A_1(V_1(\hat{z}_B))} < 1. \tag{3.158}$$

Furthermore, since $A_1(S_1(z))$ and z intersect in $\hat{z}_H$,

$$\left.\frac{dA_1(S_1(z))}{dz}\right|_{z=\hat{z}_H} > 1. \tag{3.159}$$

This inequality combined with (3.158) and the fact that $dA_1(S_1(z))/dz$ is a strictly increasing function for the assumed pgf's - for z positive real and in-

side the region of convergence of the corresponding series - it follows that

$$\hat{z}_B A_1(V_1(\hat{z}_B)) < \hat{z}_H. \tag{3.160}$$

Since $A_1(V(\hat{z}_B)) > 1$ it follows from the previous inequality that $\hat{z}_B < \hat{z}_H$, and thus $\hat{z}_H$ is never a dominant singularity of $D(z)$. It is also intuitively clear that - because of the priority scheduling discipline - the behavior of $d(n)$ for high n will be dominated by the delay of the class-2 packets.

From expression (3.129) and the fact that the dominant singularities of $D(z)$ and $D_2(z)$ are equal, it is easily seen that

$$d(n) \triangleq \text{Prob}[d = n] \tag{3.161}$$

$$\approx \frac{\lambda_2}{\lambda_T} d_2(n), \tag{3.162}$$

for large enough n. Since $d_2(n)$ is approximately calculated in expression (3.155), $d(n)$ is approximately determined from this expression. Finally, the probability that the steady-state delay of a random packet is larger than a bound D is given by

$$\text{Prob}[d > D] \approx \frac{\lambda_2}{\lambda_T} \text{Prob}[d_2 > D], \tag{3.163}$$

for large D.

3.7 Waiting time

The waiting time of a packet, defined as the number of slots a packet has to wait in the *queue* before starting service, is easily determined using the results obtained in the previous section for the packet delay. Indeed, since the service times of the packets are not interrupted, the waiting time of a packet equals its delay minus its service time. Thus, the pgf of the steady-state waiting time of a class-j packet is given by

$$T_j(z) = \frac{D_j(z)}{S_j(z)}, \tag{3.164}$$

for $j = 1, 2$. Using expressions (3.115) and (3.128), we find

$$T_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{z-1}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1} \left(\frac{1-\rho_T}{1-\rho_1} + \frac{\rho_2}{1-\rho_1} \frac{S_2(z)-1}{\mu_2(z-1)} \right), \tag{3.165}$$

for the pgf of the class-1 waiting time and

$$T_2(z) = \frac{1-\rho_T}{\lambda_2} \frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))} \frac{zA_1(V_1(z))-1}{V_2(z)-1}, \tag{3.166}$$

for the pgf of the class-2 waiting time. Finally, the pgf of the steady-state waiting time of a random packet is given by

$$T(z) = \frac{\lambda_1}{\lambda_T} T_1(z) + \frac{\lambda_2}{\lambda_T} T_2(z) \tag{3.167}$$

$$= \frac{1-\rho_1}{\lambda_T} \frac{z-1}{z - A_1(S_1(z))} \frac{A_1(S_1(z))-1}{S_1(z)-1} \left(\frac{1-\rho_T}{1-\rho_1} + \frac{\rho_2}{1-\rho_1} \frac{S_2(z)-1}{\mu_2(z-1)} \right) \tag{3.168}$$

$$+ \frac{1-\rho_T}{\lambda_T} \frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))} \frac{zA_1(V_1(z))-1}{V_2(z)-1}.$$

From these pgf's, performance measures can be calculated as is done in the previous section.

3.8 Numerical examples

To conclude this chapter, we apply our results to an output queueing switch with an NP priority scheduling discipline. We will specifically focus on the performance measures of the system contents and of the packet delay (analyzed in sections 3.3 and 3.6 respectively).

Because of the rather large number of input parameters one can vary in this model, the numerical examples discussed in this section are merely a (limited) subset of the possible numerical examples. We will furthermore focus on one type of arrival distribution (see further). This gives us a chance of limiting the number of plots, without losing too much generality.

3.8.1 Input processes

The arrival process

We use the example of the output queueing switch, discussed in section 1.7.2, throughout this section. The pgf of the number of per-slot class-1 and class-2 arrivals is given by

$$A(z_1,z_2) = \left(1 - \frac{\lambda_1}{N}(1-z_1) - \frac{\lambda_2}{N}(1-z_2)\right)^N, \tag{3.169}$$

with N the number of in- and outlets of the output queueing switch and with λ_j the probability that a class-j cell arrives at a randomly chosen inlet (for more details see 1.7.2). We will briefly repeat the most important characteristics of this arrival process: the marginal pgf $A_T(z)$ is given by

$$A_T(z) = \left(1 - \frac{\lambda_T}{N}(1-z)\right)^N, \tag{3.170}$$

and the marginal pgf of the number of per-slot arrivals of class-j is given by

$$A_j(z) = \left(1 - \frac{\lambda_j}{N}(1-z)\right)^N, \tag{3.171}$$

with $j = 1, 2$. The means of the total, class-1 and class-2 numbers of per-slot arrivals are thus given by λ_T, λ_1 and λ_2 respectively, while the variances of these three stochastic variables are given by

$$\text{Var}[a] = \lambda\left(1 - \frac{\lambda}{N}\right), \tag{3.172}$$

with λ equal to λ_T, λ_1 and λ_2 respectively. The covariance of the number of per-slot class-1 and class-2 arrivals is given by

$$\text{Cov}[a_1, a_2] = -\frac{\lambda_1\lambda_2}{N}. \tag{3.173}$$

The service process

In this section, the service times of both classes are assumed deterministic in most examples, i.e.,

$$S_j(z) = z^{\mu_j}, \tag{3.174}$$

$j = 1, 2$, with μ_j the class-j service time.

In order to study the influence of the variance of the service times though (since the variance of deterministic service times is obviously equal to 0), we assume in some examples the class-j service times to be equal to $\mu_j^{(1)}$ with probability p_j and equal to $\mu_j^{(2)}$ with probability $1 - p_j$, i.e.,

$$S_j(z) = p_j z^{\mu_j^{(1)}} + (1-p_j) z^{\mu_j^{(2)}}. \tag{3.175}$$

The mean value and variance corresponding with this distribution equal

$$\mu_j = p_j\mu_j^{(1)} + (1-p_j)\mu_j^{(2)}, \tag{3.176}$$

and

$$\text{Var}[s_j] = p_j(1-p_j)\left(\mu_j^{(2)} - \mu_j^{(1)}\right)^2, \tag{3.177}$$

respectively. In order to study the influence of Var$[s_j]$, p_j, $\mu_j^{(1)}$ and $\mu_j^{(2)}$ will be varied so that μ_j stays constant and Var$[s_j]$ is varied from 0 to infinity (in discrete steps).

3.8.2 Influence of load on moments

Firstly, we show the influence of some load characteristics on the mean and variance of the system contents and packet delay. The arrival process is defined by expression (3.169) with $N = 16$ and the service times of class-j packets are deterministically equal to μ_j. We define α as the fraction of class-1 load in the overall traffic mix, i.e.,

$$\alpha \triangleq \frac{\rho_1}{\rho_T}, \tag{3.178}$$

with $\rho_j = \lambda_j \mu_j$ $(j = 1, 2)$ and $\rho_T = \rho_1 + \rho_2$ (as before).

Since we have extensively studied the influence of the load on the means and variances of the system contents and delays in subsection 2.6.2 - in the case of deterministic service times of one slot - we limit the number of examples in this subsection. However, all types of figures from subsection 2.6.2 can be 'replicated' in the context of this chapter.

System contents

In Figures 3.4 and 3.5, the mean values and variances of the system contents of class-1 and class-2 packets are shown as functions of the total load, when $\alpha = 0.25$, 0.5 and 0.75 respectively and when the service times of both classes are deterministically equal to 20 slots. We have also shown the mean value and variance of the system contents of any class (class-1 or class-2) for $\alpha = 0.5$ when a FIFO scheduling discipline is applied. These can be easily calculated because - in the special case of the arrival process characterized by (3.169) and equally distributed service times of both classes - the joint pgf of the number of arrivals of both classes has the feature that it can be expressed as $A_T(\alpha z_1 + (1-\alpha)z_2)$. The joint pgf of the system contents of both classes is thus also given by $U_T(\alpha z_1 + (1-\alpha)z_2)$, with $U_T(z)$ the pgf of the total system contents (given by expression (3.39)) - for more details see subsection 2.6.2 - and the mean and variance are thus easily obtainable from this pgf. From Figures 3.4 and 3.5, one can see the influence of the priority scheduling discipline: especially for high loads, the mean and the variance of the number of class-1 packets in the

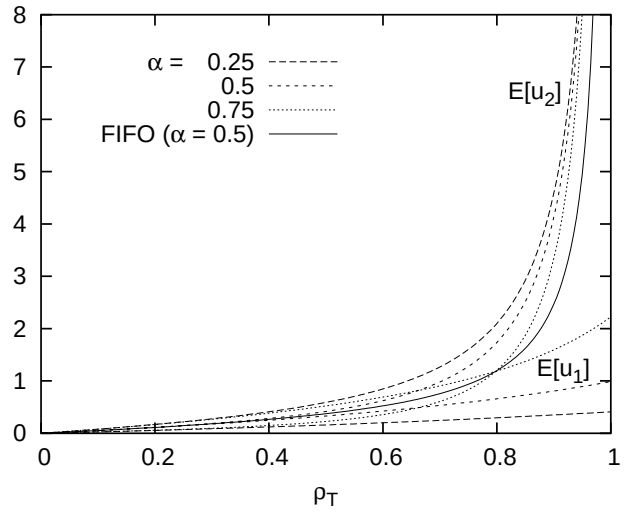


Figure 3.4: Mean values of system contents versus the total arrival rate ($\mu_1 = \mu_2 = 20$)

system are reduced by the priority scheduling discipline; the opposite holds for class-2 packets.

Packet delays

In Figures 3.6 and 3.7, the mean values and variances of the packet delay of class-1 and class-2 packets are shown as functions of the total load ρ_T, when $\mu_1 = \mu_2 = 2$ with α equal to 0.25, 0.5 and 0.75 respectively. In order to compare with a FIFO scheduling discipline, we have also shown the mean value and variance of the packet delay of any packet in that case. Because of the service times of the class-1 and class-2 packets being equally distributed and because of the specific arrival process considered in this section, the packet delay is the same for class-1 and class-2 packets in case of FIFO scheduling, and can thus be calculated as if there is only one class of packets arriving according to an arrival process with pgf $A_T(z)$. This situation has already been analyzed, e.g., in Bruneel and Kim [1993]. One can observe the influence of the priority scheduling discipline: mean value and variance of the delay of class-1 packets reduce significantly. The price to pay is a larger mean value and variance of the delay for class-2 packets. Also, for this parameter set, the smaller the fraction α of class-1 packets in the overall traffic mix, the lower the mean value and variance of the packet delay of both classes will be. This is not always the case however as can be deduced from Figure 3.8, which shows the mean delay of class-1 and class-2 packets as a function of ρ_T, when $\mu_1 = 2$ and $\mu_2 = 20$. α is again 0.25, 0.5 and 0.75 respectively. In this case, if the load is smaller

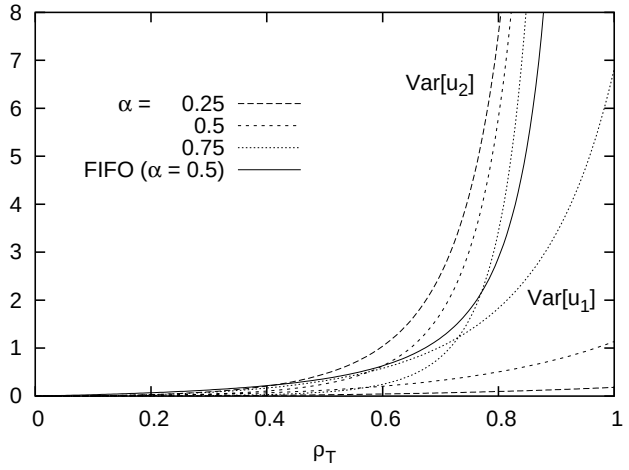


Figure 3.5: Variances of system contents versus the total arrival rate ($\mu_1 = \mu_2 = 20$)

than ≈ 0.9, the smaller the fraction of class-1 packets in the overall traffic mix, the higher the mean packet delay of both classes will be. For loads higher than 0.9, the opposite holds. This can be explained as follows. For low and moderate values of the load - due to the long service times of class-2 packets - the delay of class-1 packets will be determined by the probability of having to wait for a class-2 residual service time upon arrival and is therefore highest when the share of the class-2 packets in the overall traffic mix is highest. The same holds for the delay of the class-2 packets since there are too few class-1 packet arrivals to have a severe impact on their delay. As the total load further increases however, the transmission of class-1 packets becomes more frequent and starts to take its toll.

3.8.3 Influence of the service times on mean values

In this subsection, we show the influence of the service times on the mean system contents and mean packet delay. To show the influence of the *mean* service times of both classes, we assume deterministic service times. In order to show the influence of the *variances* of the class-1 and class-2 service times (on the mean system contents and mean packet delay) however, service times with a pgf as in (3.175) are also considered (when appropriate).

System contents

In the first two figures, we assume deterministic service times for both classes. Figure 3.9 shows the mean system contents of both classes versus the class-1

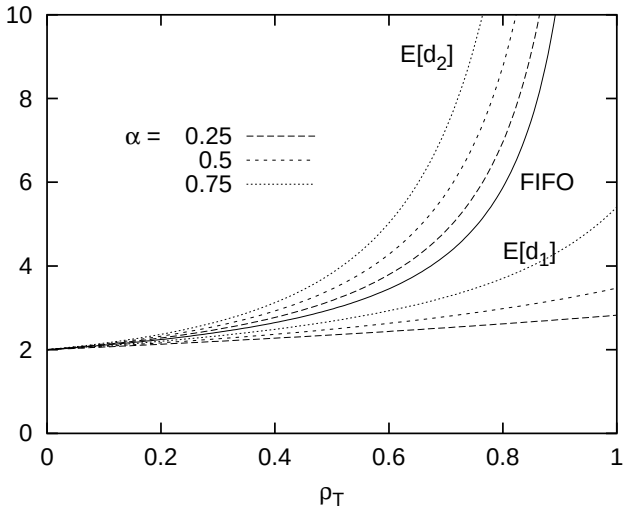


Figure 3.6: Mean values of the packet delay versus the total load ($\mu_1 = \mu_2 = 2$)

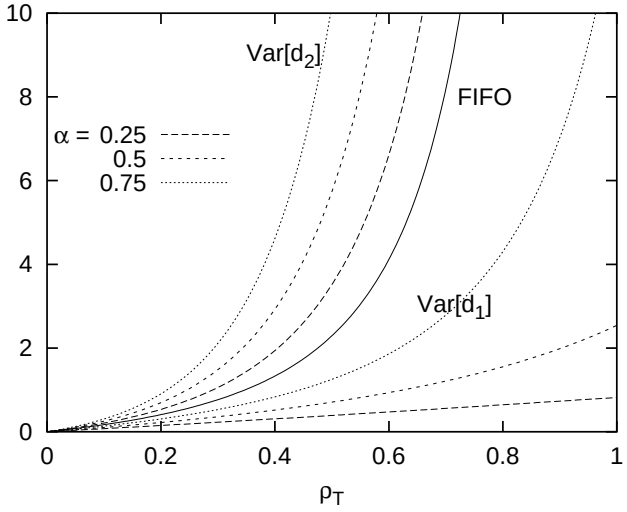


Figure 3.7: Variances of the packet delay versus the total load ($\mu_1 = \mu_2 = 2$)

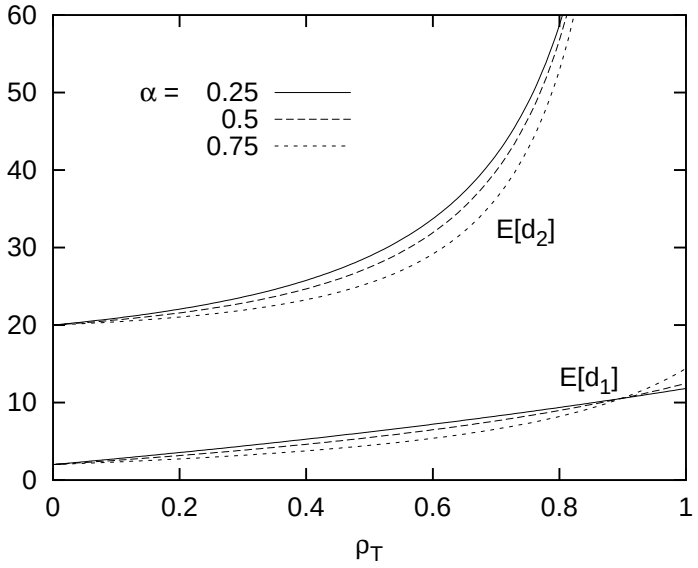


Figure 3.8: Mean values of the packet delay versus the total load ($\mu_1 = 2, \mu_2 = 20$)

service time with $\mu_2 = 20$ and $\alpha = 0.25$, 0.5 and 0.75 respectively. The total load is equal to 0.75. It can be seen that the mean class-1 system contents decreases with increasing μ_1. This is explained as follows: less class-1 packets arrive when μ_1 increases (since ρ_1 is kept constant, λ_1 decreases when μ_1 increases). The mean class-2 system contents on the other hand increases with increasing μ_1 (when $\text{Var}[s_1] = 0$). So, although the load of class-1 packets stays the same, the mean class-2 contents (heavily) increases with the class-1 service time. This is due to the fact that when the class-1 packets are longer, the variance in the lengths of the periods that the server is available and unavailable for class-2 packets grows. It is a well-known fact from queues with server interruptions (i.e., queues with a server that is unavailable for certain periods) that a higher variance in the interruption process gives rise to higher mean system contents (see e.g. Bruneel and Kim [1993], Fiems [2004]). The same phenomenon is the cause of the behavior of the mean class-2 system contents in this plot.

In Figure 3.10, the mean system contents of both classes are plotted as a function of the (mean) class-2 service time with $\mu_1 = 20$, $\rho_T = 0.75$ and $\alpha = 0.25$, 0.5 and 0.75. It can be seen that the mean class-1 contents increases, while the mean class-2 system contents decreases with increasing μ_2. The increase of the mean class-1 system contents is easily explained as follows: as seen in subsection 3.3.3, the influence of the class-2 packets on the class-1 system contents is closely related to (the number of class-1 arrivals during) the residual service times of class-2. Obviously residual service times increase when service times

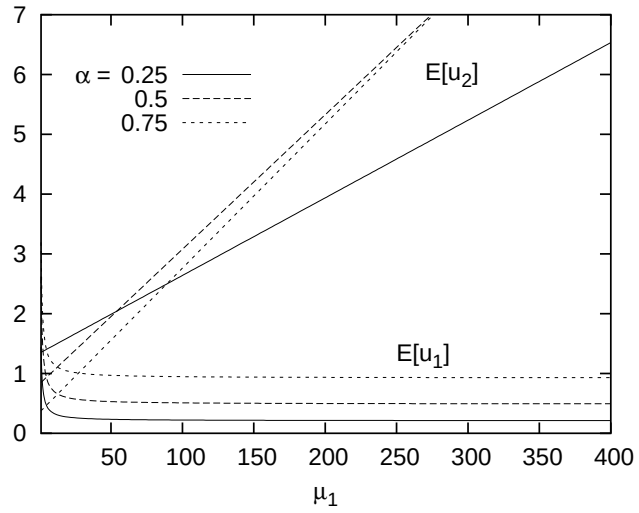


Figure 3.9: Mean values of system contents versus the (mean) class-1 service times ($\rho_T = 0.75, \mu_2 = 20$)

increase. Or in other words, when the service times of the class-2 packets increase, more class-1 packets will arrive during the time periods that class-2 packets are being served. Since the arriving packets cannot interrupt the service of these class-2 packets, they have to wait until this packet leaves the system, thus leading to a larger mean class-1 system contents. The decrease of the mean class-2 system contents is caused by the decrease of λ_2 when μ_2 increases (and ρ_2 is constant), i.e., on average less class-2 packets arrive.

As can be seen from formulas (3.59) and (3.61), the mean class-1 and class-2 system contents are linearly dependent on the variance of the class-1 and class-2 service times. This is also shown in Figures 3.11 and 3.12. In Figure 3.11 (Figure 3.12 respectively), the mean class-1 and class-2 system contents versus the variance of the class-1 (class-2 respectively) service times are depicted when the total load is 0.75, when $\alpha = 0.25$, 0.5 and 0.75, when the mean class-1 (class-2 respectively) service time is equal to 20 and when the class-2 (class-1 respectively) service times are deterministically equal to 20. The service times of class-1 (class-2 respectively) have a pgf as defined in expression (3.175), with $\mu_j^{(1)} = 1$, $\mu_j^{(2)}$ varying from 20 to infinity and p_j chosen so that the mean service time μ_j is kept constantly equal to 20 slots. It can be seen that mean class-1 and mean class-2 system contents increase with increasing $\mathrm{Var}[s_j]$. The influence on the mean class-2 system contents is generally larger than the influence on the mean class-1 system contents, due to the priority scheduling discipline.

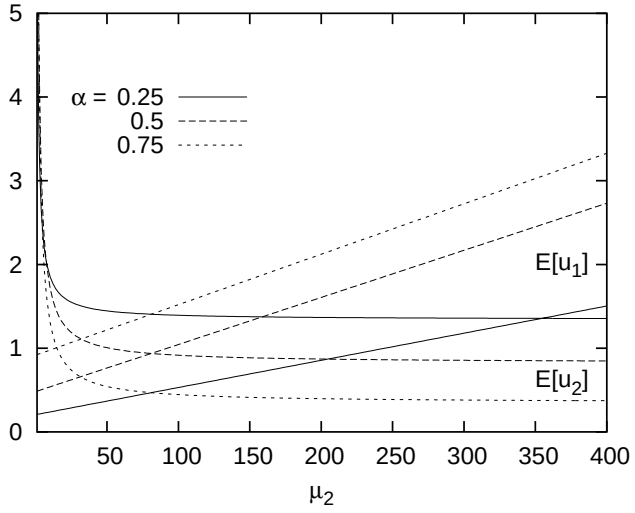


Figure 3.10: Mean values of system contents versus the (mean) class-2 service times ($\rho_T = 0.75, \mu_1 = 20$)

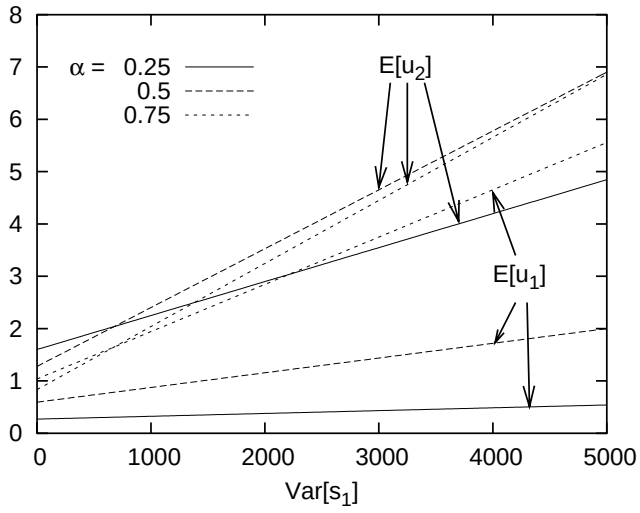


Figure 3.11: Mean values of system contents versus the variance of the class-1 service times ($\rho_T = 0.75, \mu_1 = \mu_2 = 20$)

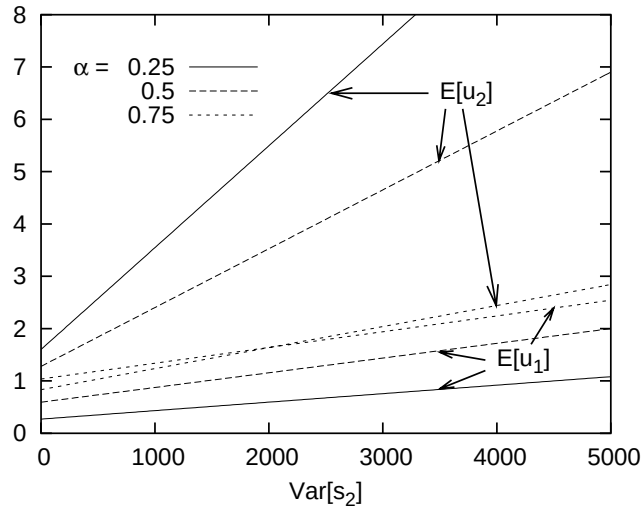


Figure 3.12: Mean values of system contents versus the variance of the class-2 service times ($\rho_T = 0.75, \mu_1 = \mu_2 = 20$)

Packet delays

Figure 3.13 shows the mean packet delays of class-1 and class-2 packets as functions of the service time of class-1 packets, when $\rho_T = 0.75$, $\mu_2 = 2$ and α is, as before, 0.25, 0.5 and 0.75. Service times of class-1 and class-2 are deterministically equal to μ_1 and μ_2 respectively. We see that the mean packet delays of both classes are proportional with μ_1 and that the impact of μ_1 on the delay of both priority classes is significant. Figure 3.14 shows the mean packet delays of class-1 and class-2 packets as functions of the service time of class-2 packets, when $\rho_T = 0.75$, $\mu_1 = 2$ and α is 0.25, 0.5 and 0.75. Varying the mean service time of class-2 packets has a considerable influence on the mean delay of class-2 packets, while the influence on the mean packet delay of class-1 packets is small, but not negligible. Furthermore we observe that for small class-2 packets, both classes have a smaller mean delay if the fraction of the load of class-1 packets is lower, while for long service times of the class-2 packets, the opposite holds, as already discussed before (see also Figure 3.8).

To emphasize the possible influence of the (mean) service times of the class-2 packets on the (mean) class-1 packet delay, we show the mean value of the packet delay of class-1 packets as a function of the total load, when $\lambda_1 = 0.25$, $\mu_1 = 2$ and $\mu_2 = 1, 2, 4, 8, 16$ in Figure 3.15. This figure shows the influence of the *NP* priority scheduling discipline. When the service time of a class-2 packet is assumed to be deterministically equal to one slot, i.e., $\mu_2 = 1$, the NP priority scheduling has the same effect - on the mean class-1 delay - as the preemptive priority scheduling, and $\mathrm{E}[d_1]$ is not influenced by the presence

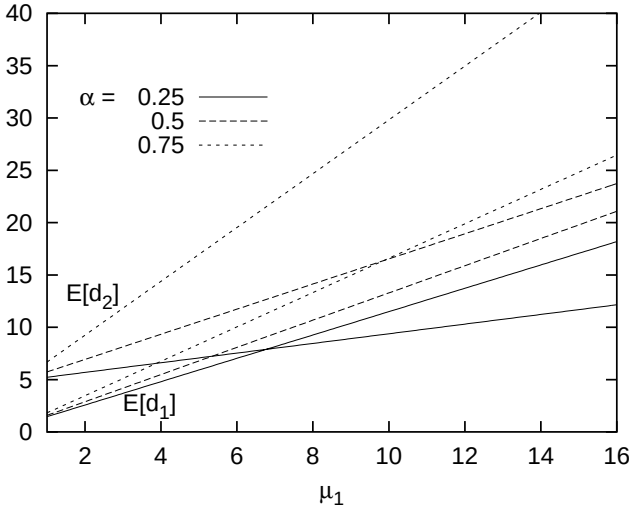


Figure 3.13: Mean packet delays versus mean service times of class-1 packets ($\rho_T = 0.75, \mu_2 = 2$)

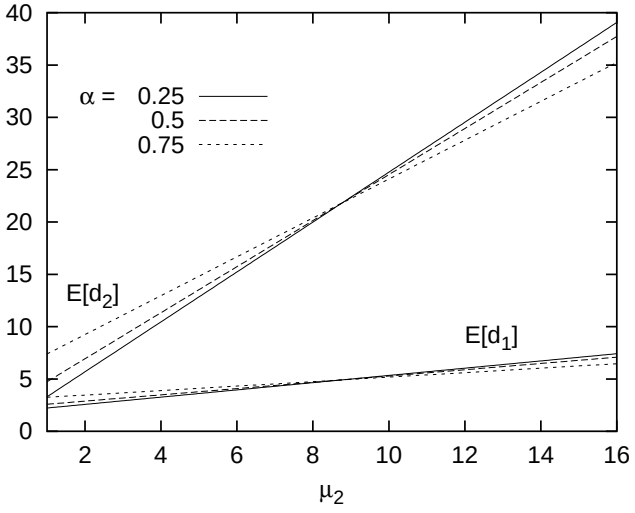


Figure 3.14: Mean packet delays versus mean service times of class-2 packets ($\rho_T = 0.75, \mu_1 = 2$)

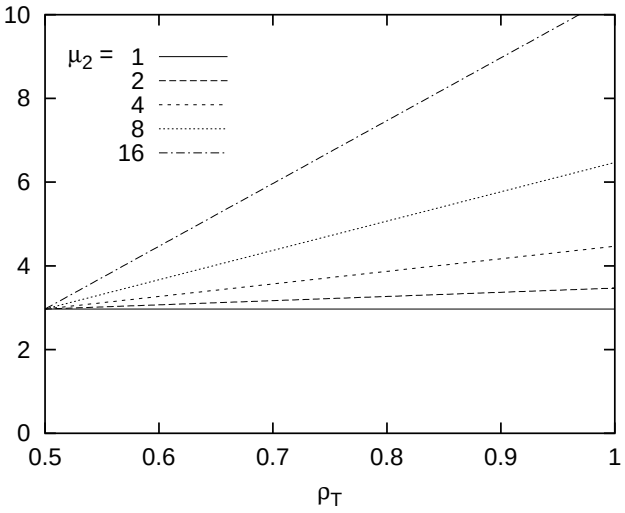


Figure 3.15: Mean packet delay of class-1 packets versus the total load ($\rho_1 = 0.5, \mu_1 = 2$)

of class-2 packets. For $\mu_2 > 1$, the higher the value of ρ_T (and hence ρ_2), the higher the probability that a newly arriving class-1 packet has to wait for a class-2 packet service completion, and obviously the effect becomes worse as μ_2 increases.

In the next two figures, we assume the service time of class-1 or class-2 packets respectively to have a pgf as defined in expression (3.175). The weight p_j is again chosen in such a way that the mean service time remains constant. By varying $\mu_j^{(2)}$ while keeping $\mu_j^{(1)}$ constant the variance of the class-j service time can be varied (from 0 until ∞). In Figure 3.16, we have plotted the mean delays of both classes as functions of the variance of the service times of class-1 packets, when $\rho_T = 0.75$, deterministic class-2 service times, $\mu_1 = \mu_2 = 2$ and $\alpha = 0.25, 0.5$ and 0.75 respectively. Figure 3.17 shows the mean delays of both classes as functions of the variance of the service times of class-2 packets, when $\rho_T = 0.75$, deterministic class-1 service times, $\mu_1 = \mu_2 = 2$ and $\alpha = 0.25, 0.5$ and 0.75 respectively. These figures illustrate that even though the mean lengths of class-1 and class-2 packets are kept constant, their variances have a large impact on both the class-1 and class-2 mean packet delays.

3.8.4 Tail probabilities

In the next figures, we illustrate the tail behavior of the packet delay. The tail behavior of the system contents is similar and as a result similar plots as the ones for the packet delay can be constructed. We have shown in section

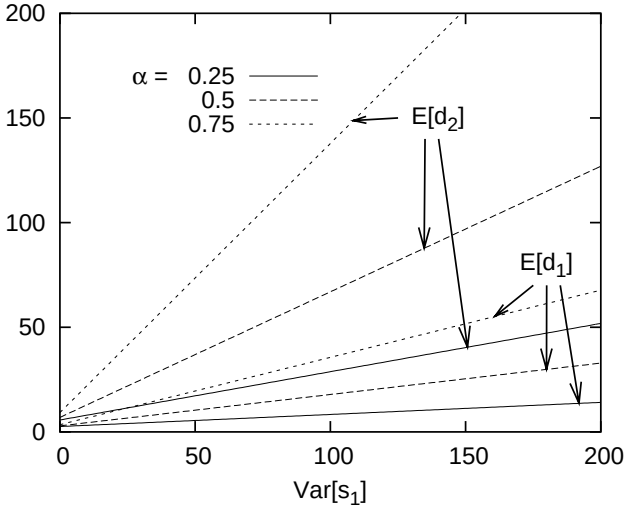


Figure 3.16: Mean values of the packet delays versus the variance of the class-1 service times ($\rho_T = 0.75, \mu_1 = \mu_2 = 0.75$)

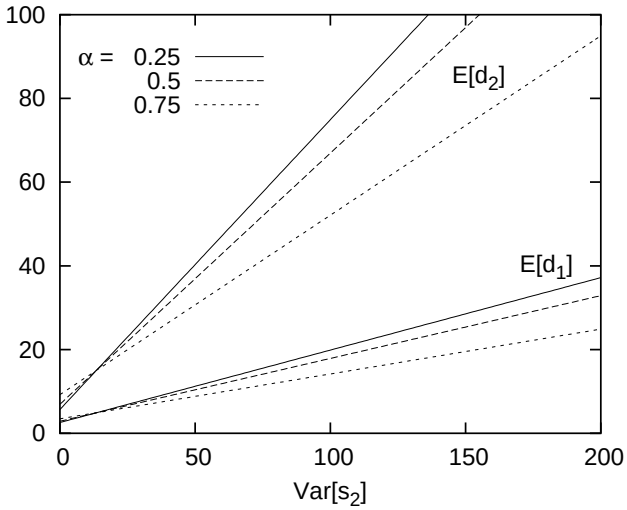


Figure 3.17: Mean values of packet delays versus the variance of the class-2 service times ($\rho_T = 0.75, \mu_1 = \mu_2 = 0.75$)

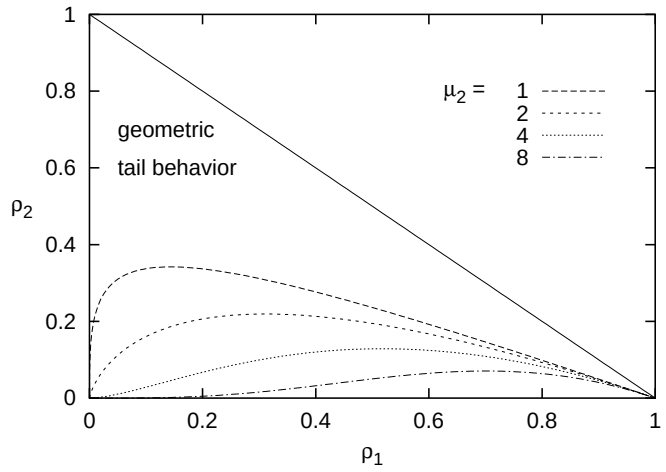


Figure 3.18: Regions for tail behavior as a function of the load of both classes in the case of deterministic class-1 service times ($\mu_1 = 2$)

3.6.6, that the tail probabilities of the class-2 packet delay can have 3 types of behavior, depending on which singularity of $D_2(z)$ is dominant. In case of the arrival process considered in this section, Figures 3.18 and 3.19 show for which combination of class-1 and class-2 loads the transition type behavior occurs for the packet delay when $\mu_1 = 2$ and for several values of μ_2, i.e., for which combination of loads the regular pole and the branch point coincide. The service times of the class-2 packets are deterministic. In Figure 3.18, the service times of the class-1 packets are also assumed to be deterministic, while in Figure 3.19 the service times of class-1 packets are geometrically distributed. In the region above the curves, the tail behavior is geometric for the respective ρ_1 and ρ_2, while below the curves the tail behavior is non-geometric. Note that in the area above the line defined by $\rho_1 + \rho_2 = 1$ in both figures, the total load is larger than 1, and as a result, the system becomes unstable. As can be seen from these figures, the higher the mean service time of class-2 packets, the smaller the region where the tail behavior is non-geometric. By comparing both figures (and from other extensive examples), we conclude that the transition between geometric and non-geometric tails highly depends on the service time distribution of the class-1 packets.

Figure 3.20 shows the tail probabilities of the packet delay of class-1 and class-2 packets for deterministic service times ($\mu_1 = \mu_2 = 2$), if $\rho_1 = 0.4$ and $\rho_2 = 0.1$ (non-geometric behavior), approximately 0.21 (transition type behavior) and 0.4 (geometric behavior) respectively. Tail behavior of the packet delay of class-1 packets is not the same for the 3 cases, but the curves lie (in this case) near to each-other. We have also compared our approximations with simulation results (marks in the figures). The figure shows that the approximations

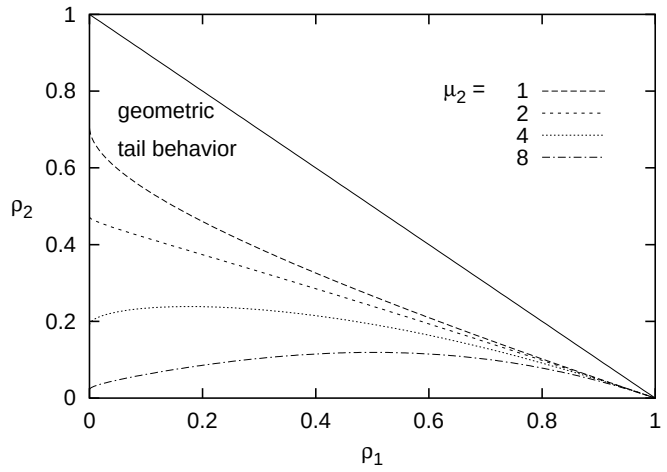


Figure 3.19: Regions for tail behavior as a function of the load of both classes in the case of geometric class-1 service times ($\mu_1 = 2$)

for the tail probabilities of the delay of both classes are very good.

Finally, Figure 3.21 shows the tail probabilities of the class-1 delay for $\rho_1 = \rho_2 = 0.4$, $\mu_1 = 2$ and several values of μ_2 ($\mu_2 = 1, 2, 4, 8, 16$). The service times of both classes are deterministic. This figure clearly shows that the class-2 service times can have a big effect on the tail probabilities of the class-1 delay. Since class-1 traffic is delay-sensitive, this is something to keep in mind when incorporating priority scheduling disciplines (e.g., in telecommunication networks).

3.9 Concluding remarks

In this chapter, we studied a more evolved priority queueing system with generally distributed service times. The priority scheduling was of the NP type. The analysis was an extension of techniques used in the previous chapter. The most important step was finding an embedded Markov chain. This was done by first analyzing the system contents at specific time instants. The two-dimensional pgf of the steady-state class-1 and class-2 system contents at the beginning of such slots was the starting-point of all further calculations in this chapter.

An important conclusion is that the NP priority discipline can differentiate the quality of service of different classes. More precisely it is shown that the delay of the high-priority traffic can be decreased by applying the NP priority discipline. The price to pay is an increase in the delay of the low-priority

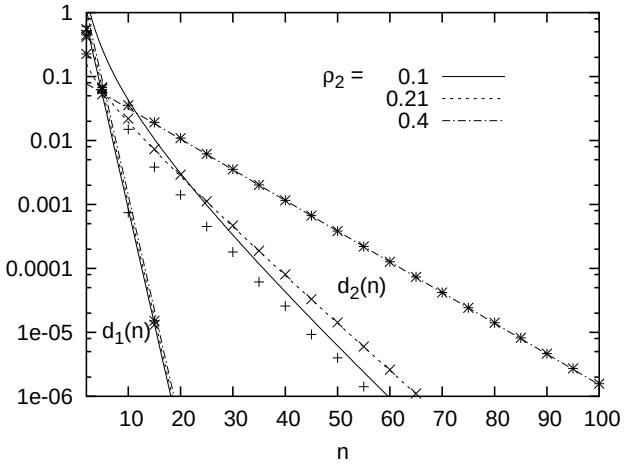


Figure 3.20: Tail behavior of the class-1 and class-2 packet delay for several class-2 loads ($\rho_1 = 0.4, \mu_1 = \mu_2 = 2$)

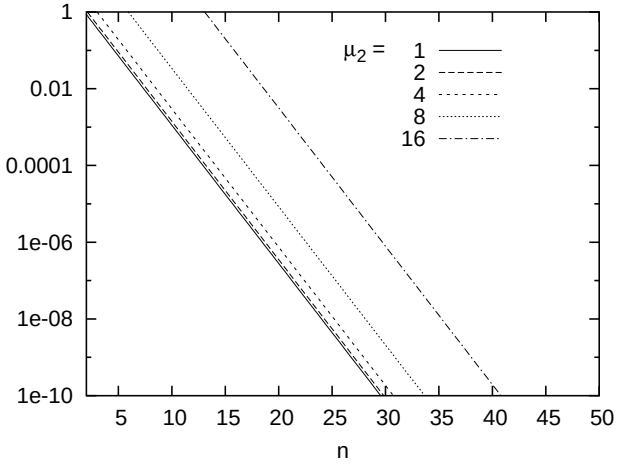


Figure 3.21: Tail behavior of the class-1 packet delay for several class-2 service times ($\rho_1 = \rho_2 = 0.4, \mu_1 = 2$)

traffic. However, it is also shown that the NP property of this scheduling discipline can have an important 'negative' influence on the high-priority delay characteristics. Especially when the service times of the low-priority class are relatively large (compared to the service times of the high-priority packets), the *positive* influence of the priority discipline on the high-priority delay can be severely *decreased* by the NP property.

Chapter 4

Preemptive resume priority

In this chapter, we describe the analysis of a *preemptive resume* (PR) priority queue. Packets of two classes (class-1 and class-2) arrive in a single-server queueing system where packets of class-1 have (PR) priority over class-2 packets. Whenever the server becomes available, a class-1 packet is served next (if any). If no class-1 packets are present, a class-2 packet can start service. In contrast with the NP (non-preemptive) priority scheduling discipline analyzed in the previous chapter, newly arriving class-1 packets interrupt an on-going service of a class-2 packet. So, when class-1 packets arrive during a service slot of a class-2 packet, one of the class-1 packets starts service at the beginning of the next slot, and the (not-yet-served part of the) interrupted class-2 packet is pushed back in the queue (unless it was its last slot of service). In this chapter, we analyze the preemptive *resume* priority scheduling discipline. An interrupted class-2 service resumes after all class-1 packets have left the system. Or in other words, only the not-yet-served part of the interrupted packet has to be served afterwards.

According to Miller [1960], the first published results for the preemptive (resume) priority queue are by White and Christie [1958]. In [Miller 1960], the Laplace-Stieltjes transform of the waiting time is found for a continuous-time PR priority queue with Poisson arrivals and generally distributed service times. Furthermore - as already mentioned in the previous chapter - an overview of some other basic (non)-preemptive priority queueing models in continuous-time can be found in the monographs of Kleinrock [1976] and Takagi [1991].

Continuous-time single-server PR priority queues with no correlation between the arrival processes of the priority classes have been analyzed by a.o. Marks [1973], Miller [1981], Brandwajn [1982], Cidon and Sidi [1990], Takahashi and Miyazawa [1994], Takine and Hasegawa [1994], Abate and Whitt [1997], Boxma et al. [1999], Berger and Whitt [2000], Kraimeche [2001], Drekic and Grassmann [2002], Drekic and Stafford [2002], Sharma and Virtamo [2002],

Takada and Miyazawa [2002] and Liu and Gong [2003]. Marks [1973] gives an algorithm for the calculation of the state probabilities in a PR priority queue with negative exponential interarrival and service times. Miller [1981] analyzes a PR priority queue with Poisson arrivals and exponential service times. The state probabilities of the number of packets of each type in the system are presented by explicit recursive formulas. Brandwajn [1982] demonstrates a technique whereby one attempts to solve the difference equations - the starting point of most queueing problems - directly for PR priority queues with rather general input processes. A two-class PR priority queue with Poisson arrivals and exponential service times is studied by Drekic and Grassmann [2002]. The buffer storing the high-priority packets is assumed to be infinitely large, while the low-priority buffer is of finite length. The method of the generalized eigenvalues is used to determine the joint distribution of the number of high- and low-priority packets. A recursive formula for the steady-state probabilities of the system contents of the two priority classes is proposed by Cidon and Sidi [1990]. This formula is found from the joint pgf of these stochastic variables, as calculated by Miller [1960]. Abate and Whitt [1997] quantify the effect of the priority structure on the low-priority steady-state delay tail probabilities (in NP and PR priority systems). They show that the priority structure tends to make the tail probabilities non-exponential (as also shown in this dissertation). Takahashi and Miyazawa [1994] propose a relationship between the queue contents and the waiting time for the NP and PR priority queues. Takine and Hasegawa [1994] study a queue with Markovian arrivals and with state-dependent service time distributions. As an application of the results obtained for this queueing model, a PR priority queue is analyzed. In particular, the Laplace-Stieltjes transform and the mean values of the waiting times of packets of all priority classes are obtained. Boxma et al. [1999] derive a heavy-traffic limit theorem for the low-priority waiting time when the service times are heavy-tailed. Upper and lower bounds on the per-class workload distributions in a priority fluid queue are established by Berger and Whitt [2000]. Kraimeche [2001] models the multiplexing of video and data sources in an ATM access network by a fluid priority queue. The cell loss of the data traffic is analyzed. Drekic and Stafford [2002] propose a symbolic computation procedure to calculate (higher order) moments of the system contents and packet delay - starting from the pgf or Laplace-Stieltjes transform of these variables - in (general) priority queues. Sharma and Virtamo [2002] study a two-class PR priority queue with (Markov modulated) Poisson arrivals and general service times. The queue is assumed to be of finite length in the sense that the total amount of (unfinished) work in the queue is assumed to be limited (in contrast with traditional finite queue analyses, where the number of packets in the system is assumed to be limited). Takada and Miyazawa [2002] analyzed a fluid priority queue with the high-priority traffic existing out of continuous fluid and the low-priority traffic out of batch fluid. The stationary joint distribution of the buffer contents of both types is obtained in terms of matrix transforms. In Liu and Gong [2003], a two-class

fluid priority queue is studied with the high-priority traffic modeled by the superposition of on-off sources and the low-priority traffic by a constant bit rate flow. Sample path analysis tools are used to obtain various analytical results (e.g., the low-priority system contents distribution).

Continuous-time multi-server PR priority queues are analyzed by a.o. Goldberg [1981], Mitrani and King [1981], Buzen and Bondi [1983], Kao and Narayanan [1991], Gail et al. [1992], Tabet-Aouel and Kouvatsos [1992], Kouvatsos and Tabet-Aouel [1994], Núñez Queija and Boxma [1998] and van der Heijden et al. [2004]. All these papers except for Gail et al. [1992], Tabet-Aouel and Kouvatsos [1992] assume Poisson arrivals and exponential service times. In [Goldberg 1981], an exact and a (faster) approximate procedure to calculate the steady-state probabilities numerically are established. Mitrani and King [1981], Buzen and Bondi [1983] calculate the mean values of the steady-state delay. Both papers provide an exact analysis for a two-class system and an approximative analysis for a system with a general number of priority classes. Kao and Narayanan [1991] analyze a two-class PR priority queue by making use of matrix-analytic solution techniques. To limit the state space, two approximations are proposed: assuming a finite high-priority buffer or a finite low-priority buffer respectively. Gail et al. [1992] analyze a two-class PR priority queue. By eliminating remaining unknown variables from the generating functions of the state probabilities, these latter ones are numerically calculated. An entropy maximization approach is used to characterize the distributional form of the steady-state probabilities in a PR priority queue with general interarrival and service times by Tabet-Aouel and Kouvatsos [1992] and Kouvatsos and Tabet-Aouel [1994]. Núñez Queija and Boxma [1998] study a two-class priority system with independent Poisson arrivals and exponential service times. The low-priority packets are furthermore served in a processor sharing fashion. A complete characterization of the joint distribution of the steady-state system contents of both classes is given and the results are applied in an ATM context. The approximation of performance measures in multi-class PR priority queues is discussed by van der Heijden et al. [2004] for large problem instances (many classes and servers) using class aggregation and server reduction.

Discrete-time PR priority queues with *no correlation between the arrival processes of different classes* are studied by a.o. Rubin and Tsai [1989], Chen and Guérin [1991], Choi et al. [1997], Lee et al. [1998], Lee [2001] and Fiems et al. [2004]. Chen and Guérin [1991], Choi et al. [1997] and Lee et al. [1998] analyze an input queueing ATM switch with cells of two traffic classes arriving to the inlets of the switch. The cells arrive at the inlets according to a Bernoulli process and are all of the same length. Because of the head-of-the-line blocking of the input queues, the service times of cells are not deterministic but are - approximatively - modeled by a geometric distribution. Fiems et al. [2004] describe an analysis of a queue with service interruptions. The service interruptions are of renewal type. The results are used to analyze the low-priority characteristics (such as the system contents and packet delay) in preemptive

priority queues. Rubin and Tsai [1989] and Lee [2001] analyze a PR priority queue with general service times. The numbers of per-slot arrivals are i.i.d. from slot-to-slot. The mean delays of a general number of priority classes are calculated using pgf's in [Rubin and Tsai 1989]. Lee [2001] calculates the joint pgf of the system contents of two priority classes, as well as the pgf's of the unfinished work and the delay of both classes. The queueing model and the analysis method described in this chapter is closely related to the ones in this paper. The main difference is that we allow correlation between the number of per-slot arrivals in this chapter, while this is assumed not to be the case in the model in [Lee 2001].

Takahashi and Hashida [1991] and Walraevens et al. [2000a, 2001, 2002b, 2004b] study *discrete-time* PR priority queues *with correlation between the number of per-slot arrivals of the different priority classes*. The numbers of per-slot arrivals are assumed to be i.i.d. in all these papers. Takahashi and Hashida [1991] calculate the pgf of the delay of a general number of priority classes, starting from the total unfinished work (without analyzing the system contents). In [Walraevens et al. 2000a, 2001, 2002b, 2004b], the number of priority classes is assumed to be equal to two. In these papers the joint pgf of the system contents and the pgf's of the packet delays of both classes are determined. In [Walraevens et al. 2000a, 2004b], the service times are assumed to be geometrically distributed. This condition is first relaxed to general service times for the high-priority service times in [Walraevens et al. 2001]. Finally, in [Walraevens et al. 2002b], the service times of both classes are generally distributed. Note that in the models of all these papers the distributions of the service times of different classes may be different.

In this chapter, we describe the analysis of a PR priority queue with two priority classes as described in [Walraevens et al. 2000a, 2001, 2002b, 2004a,b]. First, we describe the analysis approach in section 4.1. The multivariate pgf that is the starting-point of our further calculations is described in section 4.2. The system contents, queue contents and unfinished work of both classes are analyzed in section 4.3, 4.4 and 4.5 respectively. The packet delay and waiting time are discussed in sections 4.6 and 4.7. In section 4.8, we briefly focus on some relationships with results of the NP priority queue. Some numerical examples are shown in section 4.9, before completing this chapter with some concluding remarks in section 4.10.

4.1 Preliminaries

As described in section 3.1, the system contents at consecutive slot boundaries of queues with generally distributed service times do not form a Markov chain. Therefore, a Markov chain first has to be constructed.

The embedded Markov chain technique used in the previous chapter - defining specific slot boundaries so that the system contents at these slot bound-

aries form a Markov chain - is more complicated for the PR priority queue. This is caused by the fact that the service of low-priority packets can be interrupted and later resumed. In the previous chapter for instance, we defined the start-slots as the slots at the beginning of which a packet can enter the server (if there are any packets). Because of the non-preemptive property, the number of slots between a start-slot and the next start-slot was either one, the length of a class-1 service time or the length of a class-2 service time. In the PR priority case however, this is no longer true, since class-2 service times are interrupted by newly arriving class-1 packets. It is furthermore complicated to define other time epochs so that the system contents at these time epochs form a Markov chain.

Therefore, we follow another approach in the case of queues with a preemptive (resume) priority scheduling discipline. This is a technique that is widely used in queueing theory. We construct a Markov chain using stochastic variables at consecutive slot boundaries (we thus do not define specific slot boundaries). When the system contents at consecutive slots do not form a Markov chain, *supplementary variables* are defined in such a way that the system contents together with these supplementary variables form a Markov chain. This is called the *supplementary variable technique* (and is as one of the first times used by Cox [1955]). In this chapter, we start with a relatively simple model and extend the model and analysis in a few steps.

Firstly, we assume that the service times of both classes are (shifted) geometrically distributed. The geometric distribution contains the memoryless property. In this case the memoryless property means that when tagging a slot wherein a service time is on-going the residual part of the service time from that slot on is independent of the amount of service the packet has already received. Or in other words, the probability that a packet that is served during a slot will need at least another slot of service is equal to a constant parameter β, independent of the amount of service it has already received before that slot. Therefore, the system contents at consecutive slot boundaries form a Markov chain and no supplementary variables have to be defined. Note that we allow this parameter β to be different for class-1 and class-2 service times. This model is analyzed in [Walraevens et al. 2000a, 2004b].

Secondly, we assume general service times for the class-1 packets, while keeping the service times of the class-2 packets geometrically distributed. In this case, the system contents at consecutive slot boundaries do no longer form a Markov chain. We therefore define a supplementary variable, notably, the remaining number of slots that a class-1 packet in service needs before leaving the system. This is called the *residual service time of class-1*. The system contents of both classes and the residual service time of class-1 at the beginning of consecutive slots form a Markov chain. Note, that since the class-2 service times are geometrically distributed, the class-2 system contents is still enough to describe the class-2 characteristics of the queue. This model is studied in [Walraevens et al. 2001, 2004a].

Finally, we relax the last limitation in our model and analyze the system contents in the case of generally distributed service times for both classes. In this case, we define a second supplementary variable - besides the residual service time of class-1 - notably, the *residual service time of class-2*. This stochastic variable is defined as the number of slots of service that the oldest class-2 packet in the system needs to be completely served. The oldest class-j packet in the system at a certain time instant is defined as the class-j packet that - from all the packets present in the system at that time instant - arrived first. Note, that the definition of the residual service time of class-2 is a bit different than for the class-1 residual service time. This is because the oldest class-2 packet is not necessarily in service - because of the possible presence of class-1 packets in the system which are served with priority over the class-2 packets - as opposed to the oldest class-1 packet. The analysis described for this model is as in [Walraevens et al. 2002b].

4.2 The supplementary variable technique

4.2.1 Geometrically distributed service times

First, we assume geometric service times for both classes. The pgf of the service times of class-j packets is given by

$$S_j(z) = \frac{(1-\beta_j)z}{1-\beta_j z}, \tag{4.1}$$

with $j = 1, 2$. Thus β_j denotes the probability that an on-going class-j service time lasts at least another slot.

We denote the system contents of class-j packets at the beginning of slot k by $u_{j,k}$ ($j = 1, 2$). Their joint pgf is given by

$$U_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\right]. \tag{4.2}$$

As already mentioned in section 4.1, the set $\{(u_{1,k}, u_{2,k}), k \geq 1\}$ forms a Markov chain, since the arrival process is i.i.d. and the service times are geometrically distributed. The following system equations are established:

1. If $u_{1,k} = u_{2,k} = 0$:

$$u_{j,k+1} = a_{j,k}, \tag{4.3}$$

   $j = 1, 2$, i.e., the only packets present in the system at the beginning of slot $k+1$ are the packets that arrived during slot k.

2. If $u_{1,k} = 0, u_{2,k} > 0$:

$$u_{1,k+1} = a_{1,k} \tag{4.4}$$

$$u_{2,k+1} = \begin{cases} u_{2,k} + a_{2,k} & \text{with probability } \beta_2 \\ u_{2,k} + a_{2,k} - 1 & \text{with probability } 1 - \beta_2, \end{cases} \tag{4.5}$$

   i.e., the class-2 packet in service stays in the system (not necessarily in the server) with probability β_2 and leaves the system at the end of slot k with probability $1 - \beta_2$.

3. If $u_{1,k} > 0$:

$$u_{1,k+1} = \begin{cases} u_{1,k} + a_{1,k} & \text{with probability } \beta_1 \\ u_{1,k} + a_{1,k} - 1 & \text{with probability } 1 - \beta_1 \end{cases} \tag{4.6}$$

$$u_{2,k+1} = u_{2,k} + a_{2,k}, \tag{4.7}$$

   i.e., the class-1 packet in service stays in the system with probability β_1 and leaves the system at the end of slot k with probability $1 - \beta_1$.

Using these system equations, we derive a relation between $U_k(z_1, z_2)$ and $U_{k+1}(z_1, z_2)$. We remind the reader that $\mathrm{E}[X\{Y\}]$ is short for $\mathrm{E}[X|Y]\mathrm{Prob}[Y]$. We proceed as follows, taking into account the statistical independence of the random variables $(u_{1,k}, u_{2,k})$ and $(a_{1,k}, a_{2,k})$:

$$U_{k+1}(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k+1}} z_2^{u_{2,k+1}}\right] \tag{4.8}$$

$$= \mathrm{E}\left[z_1^{a_{1,k}} z_2^{a_{2,k}} \{u_{1,k} = u_{2,k} = 0\}\right] \tag{4.9}$$

$$+ \beta_2 \mathrm{E}\left[z_1^{a_{1,k}} z_2^{u_{2,k}+a_{2,k}} \{u_{1,k} = 0, u_{2,k} > 0\}\right]$$

$$+ (1 - \beta_2)\mathrm{E}\left[z_1^{a_{1,k}} z_2^{u_{2,k}+a_{2,k}-1} \{u_{1,k} = 0, u_{2,k} > 0\}\right]$$

$$+ \beta_1 \mathrm{E}\left[z_1^{u_{1,k}+a_{1,k}} z_2^{u_{2,k}+a_{2,k}} \{u_{1,k} > 0\}\right]$$

$$+ (1 - \beta_1)\mathrm{E}\left[z_1^{u_{1,k}+a_{1,k}-1} z_2^{u_{2,k}+a_{2,k}} \{u_{1,k} > 0\}\right]$$

$$= A(z_1, z_2)\mathrm{Prob}\left[u_{1,k} = u_{2,k} = 0\right] \tag{4.10}$$

$$+ A(z_1, z_2)\left(\beta_2 + \frac{1 - \beta_2}{z_2}\right)\mathrm{E}\left[z_2^{u_{2,k}} \{u_{1,k} = 0, u_{2,k} > 0\}\right]$$

$$+ A(z_1, z_2)\left(\beta_1 + \frac{1 - \beta_1}{z_1}\right)\mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}} \{u_{1,k} > 0\}\right]$$

$$= \frac{A(z_1, z_2)}{z_1 z_2}[z_1 z_2 U_k(0, 0) + z_1(1 - \beta_2 + \beta_2 z_2)(U_k(0, z_2) - U_k(0, 0)) \tag{4.11}$$

$$+ (1 - \beta_1 + \beta_1 z_1) z_2 (U_k(z_1, z_2) - U_k(0, z_2))].$$

We assume the system is stable and as a result $U_k(z_1, z_2)$ and $U_{k+1}(z_1, z_2)$ converge both to a common steady-state value $U(z_1, z_2)$. By taking the $k \to \infty$ limit of equation (4.11), we obtain:

$$U(z_1, z_2) = \frac{A(z_1, z_2)\left\{\begin{array}{l} z_1(1-\beta_2)(z_2-1)U(0,0) \\ +[(1-\beta_2)z_1 - (1-\beta_1)z_2 + (\beta_2-\beta_1)z_1 z_2]U(0, z_2) \end{array}\right\}}{z_2[(1-\beta_1 A(z_1, z_2))z_1 - (1-\beta_1)A(z_1, z_2)]}. \tag{4.12}$$

It now remains for us to determine the unknown function $U(0, z_2)$ and the unknown parameter $U(0, 0)$. This can be done in two steps. Firstly, we notice that the joint pgf $U(z_1, z_2)$ must be bounded for all values of z_1 and z_2 such that $|z_1| < 1$ and $|z_2| < 1$. In particular, this should be true for $z_1 = Y_1(z_2)$ - with

$$Y_1(z) \triangleq E_1(Y_1(z), z) \tag{4.13}$$

$$= \frac{(1-\beta_1)A(Y_1(z), z)}{1-\beta_1 A(Y_1(z), z)}, \tag{4.14}$$

and $E_j(z_1, z_2) \triangleq S_j(A(z_1, z_2))$ - and for $|z_2| < 1$ (see the appendix for more details). The above implies that if we choose $z_1 = Y_1(z_2)$ in equation (4.12), with $|z_2| < 1$, the denominator of the right-hand side of this equation becomes zero. The same must then be true for its numerator, yielding

$$U(0, z_2) = \frac{U(0,0)Y_1(z_2)(1-\beta_2)(z_2-1)}{(1-\beta_1)z_2 - (1-\beta_2)Y_1(z_2) - (\beta_2-\beta_1)Y_1(z_2)z_2}. \tag{4.15}$$

The following expression for $U(z_1, z_2)$ is then derived from equation (4.12) together with equation (4.15):

$$U(z_1, z_2) = \frac{U(0,0)E_1(z_1, z_2)(z_1 - Y_1(z_2))}{z_1 - E_1(z_1, z_2)} \times \frac{(1-\beta_2)(z_2-1)}{(1-\beta_2)Y_1(z_2)(z_2-1) - (1-\beta_1)z_2(Y_1(z_2)-1)}. \tag{4.16}$$

Finally, in order to find an expression for $U(0, 0)$, we put $z_1 = z_2 = 1$ and use de l'Hôpital's rule in expression (4.16). Therefore, we will need to calculate the value of $Y_1'(1)$. By taking the derivative of both sides of the definition of $Y_1(z)$ and by substituting z by 1, we obtain - $Y_1(1) = 1$ since we have already proved in the previous chapter that $Y_1(z)$ is a pgf -

$$Y_1'(1) = \frac{\lambda_2 \mu_1}{1-\rho_1} \tag{4.17}$$

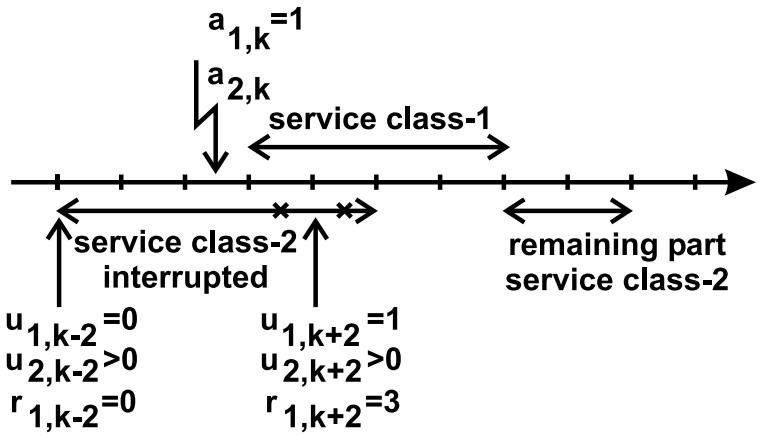


Figure 4.1: Sample of the time-axis for general class-1 and geometric class-2 service times

$$= \frac{\lambda_2}{1 - \beta_1 - \lambda_1}. \tag{4.18}$$

Using this expression, we obtain the expected result for $U(0,0)$:

$$U(0,0) = 1 - \rho_T. \tag{4.19}$$

Substituting expression (4.19) in expression (4.16) finally gives $U(z_1, z_2)$ in terms of the system's parameters and the implicitly defined $Y_1(z)$.

In the special case that $\beta_1 = \beta_2 = 0$, the service times are deterministically equal to one slot. Substituting $\beta_1 = \beta_2 = 0$ indeed yields expression (2.13) for $U(z_1, z_2)$ in chapter 2.

4.2.2 Generally distributed class-1 service times, geometrically distributed class-2 service times

In this subsection, we analyze the steady-state system contents in the case of generally distributed class-1 service times and geometrically distributed (with parameter β_2) class-2 service times. Since the service times of class-1 packets are generally distributed, the set $\{(u_{1,k}, u_{2,k}), k \geq 1\}$ does (generally) not form a Markov chain. Therefore, we introduce a new stochastic variable $r_{1,k}$ as follows: $r_{1,k}$ indicates the remaining number of slots needed to transmit the class-1 packet in service at the beginning of slot k, if $u_{1,k} > 0$, and $r_{1,k} = 0$ if $u_{1,k} = 0$. A sample of the time axis is given in Figure 4.1, in order to demonstrate the relevant stochastic variables.

With this definition, $\{(r_{1,k}, u_{1,k}, u_{2,k}), k \geq 1\}$ is seen to constitute a Markovian state description of the system at the beginning of slot k. Let $s^*_{1,k}$ indicate the service time of the next class-1 packet to receive service at the beginning of slot k. The following system equations are then established:

1. If $r_{1,k} = 0$ (and hence $u_{1,k} = 0$):
   (a) If $u_{2,k} = 0$:

$$u_{j,k+1} = a_{j,k} \tag{4.20}$$

$$r_{1,k+1} = \begin{cases} 0 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0, \end{cases} \tag{4.21}$$

   ($j = 1, 2$), i.e., the only packets present in the system at the beginning of slot $k+1$ are the packets that have arrived during the previous slot. If there are new arrivals of class-1 packets during slot k, the remaining number of slots needed to serve the packet in service at the beginning of slot $k+1$ is equal to this packet's complete service time; otherwise, a class-2 packet (if any) enters the server at the beginning of slot $k+1$.

   (b) If $u_{2,k} > 0$:

$$u_{1,k+1} = a_{1,k} \tag{4.22}$$

$$u_{2,k+1} = \begin{cases} u_{2,k} + a_{2,k} & \text{with probability } \beta_2 \\ u_{2,k} - 1 + a_{2,k} & \text{with probability } 1 - \beta_2 \end{cases} \tag{4.23}$$

$$r_{1,k+1} = \begin{cases} 0 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0. \end{cases} \tag{4.24}$$

   Expression (4.23) expresses that the service of the class-2 packet in service at the beginning of slot k is completed at the end of slot k with probability $1 - \beta_2$, or that the packet stays in the system (not necessarily the server) with probability β_2. This last event combined with $a_{1,k} > 0$ represents the case where the service of the class-2 packet is interrupted by a newly arriving class-1 packet, due to the preemptive service discipline.

2. If $r_{1,k} = 1$:

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \tag{4.25}$$

$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{4.26}$$

$$r_{1,k+1} = \begin{cases} 0 & \text{if } u_{1,k} - 1 + a_{1,k} = 0 \\ s^*_{1,k} & \text{if } u_{1,k} - 1 + a_{1,k} > 0. \end{cases} \tag{4.27}$$

   Since the class-1 packet in service at the beginning of slot k needs one more slot of service (namely slot k), it leaves the system at the end of

slot k. If there are class-1 packets present in the system at the end of slot k, the service of a new class-1 packet is started.

3. If $r_{1,k} > 1$:

$$u_{j,k+1} = u_{j,k} + a_{j,k} \tag{4.28}$$
$$r_{1,k+1} = r_{1,k} - 1, \tag{4.29}$$

for $j = 1, 2$. The class-1 packet in service at the beginning of slot k stays in the server at the beginning of slot $k + 1$. Its residual service time is decreased by one.

We define $P_k(x_1, z_1, z_2)$ as the joint pgf of the state vector $(r_{1,k}, u_{1,k}, u_{2,k})$:

$$P_k(x_1, z_1, z_2) \triangleq \mathrm{E}\left[x_1^{r_{1,k}} z_1^{u_{1,k}} z_2^{u_{2,k}}\right]. \tag{4.30}$$

Using the system equations, we derive a relation between $P_k(.,.,.)$ and $P_{k+1}(.,.,.)$. We first define

$$R_k(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k}-1} z_2^{u_{2,k}} \{r_{1,k} = 1\}\right] \tag{4.31}$$
$$= \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} \mathrm{Prob}[r_{1,k} = 1, u_{1,k} = m, u_{2,k} = n] z_1^{m-1} z_2^n. \tag{4.32}$$

For further use we also note that

$$R_k(0, z_2) = \mathrm{E}\left[z_2^{u_{2,k}} \{r_{1,k} = u_{1,k} = 1\}\right]. \tag{4.33}$$

Taking into account the statistical independence of the random variables $(r_{1,k}, u_{1,k}, u_{2,k})$ and $(a_{1,k}, a_{2,k})$ respectively, we find:

$$\begin{aligned} P_{k+1}(x_1, z_1, z_2) =& A(0, z_2) P_k(0, 0, 0) + (A(z_1, z_2) - A(0, z_2)) S_1(x_1) P_k(0, 0, 0) \\ &+ A(0, z_2) \left(\beta_2 + \frac{1 - \beta_2}{z_2}\right) (P_k(0, 0, z_2) - P_k(0, 0, 0)) \\ &+ (A(z_1, z_2) - A(0, z_2)) S_1(x_1) \left(\beta_2 + \frac{1 - \beta_2}{z_2}\right) \\ &\times (P_k(0, 0, z_2) - P_k(0, 0, 0)) + A(0, z_2) R_k(0, z_2) \\ &+ A(z_1, z_2) S_1(x_1) R_k(z_1, z_2) - A(0, z_2) S_1(x_1) R_k(0, z_2) \\ &+ \frac{A(z_1, z_2)}{x_1} [P_k(x_1, z_1, z_2) - x_1 z_1 R_k(z_1, z_2) - P_k(0, 0, z_2)]. \end{aligned} \tag{4.34}$$

We assume that the system is stable and as a result $P_k(x_1, z_1, z_2)$ ($R_k(z_1, z_2)$ respectively) and $P_{k+1}(x_1, z_1, z_2)$ ($R_{k+1}(z_1, z_2)$ respectively) converge both to a common steady-state value $P(x_1, z_1, z_2)$ ($R(z_1, z_2)$ respectively). By taking the $k \to \infty$ limit of expression (4.34), we obtain:

$$\begin{aligned} P(x_1, z_1, z_2) = & \frac{1}{z_2(x_1 - A(z_1, z_2))}[(1-\beta_2)x_1(z_2-1)\{A(0, z_2)(1 - S_1(x_1)) \\ & + A(z_1, z_2)S_1(x_1)\}P(0, 0, 0) \\ & + x_1 A(0, z_2)(1 - S_1(x_1))(1 - \beta_2 + \beta_2 z_2)P(0, 0, z_2) \\ & + A(z_1, z_2)(x_1 S_1(x_1)(1 - \beta_2 + \beta_2 z_2) - z_2)P(0, 0, z_2) \\ & + x_1 z_2 A(0, z_2)(1 - S_1(x_1))R(0, z_2) \\ & + x_1 z_2 A(z_1, z_2)(S_1(x_1) - z_1)R(z_1, z_2)]. \end{aligned} \tag{4.35}$$

It now remains for us to determine the unknown functions $P(0, 0, z_2)$, $R(0, z_2)$ and $R(z_1, z_2)$ and the unknown parameter $P(0, 0, 0)$ from equation (4.35). This can be done in the following steps. Firstly, we observe that $P(x_1, 0, z_2) = P(0, 0, z_2)$ for all x_1 and z_2, due to the fact that $r_{1,k} = 0$ iff $u_{1,k} = 0$. Substituting z_1 by 0 in equation (4.35) and using this property, we obtain:

$$R(0, z_2) = \frac{\left\{ \begin{array}{l} [z_2 - A(0, z_2)(1 - \beta_2 + \beta_2 z_2)]P(0, 0, z_2) \\ -(1 - \beta_2)(z_2 - 1)A(0, z_2)P(0, 0, 0) \end{array} \right\}}{z_2 A(0, z_2)}. \tag{4.36}$$

This expression can be used to eliminate $R(0, z_2)$ in expression (4.35).

Next, we notice that the function $P(x_1, z_1, z_2)$ must be bound for all values of x_1 and z_j such that $|x_1| < 1$ and $|z_j| < 1$ ($j = 1, 2$) since $P(x_1, z_1, z_2)$ is a pgf. In particular, this should be true for $x_1 = A(z_1, z_2)$ and $|z_j| < 1$ ($j = 1, 2$), since $|A(z_1, z_2)| < 1$ for all such z_j ($A(z_1, z_2)$ is a pgf). The above implies that if we choose $x_1 = A(z_1, z_2)$ in equation (4.35), where $|z_j| < 1$, the denominator of the right-hand side of this equation vanishes. The same must then be true for its numerator, which yields the following relation for $R(z_1, z_2)$:

$$R(z_1, z_2) = \frac{E_1(z_1, z_2)\left\{ \begin{array}{l} A(z_1, z_2)(1 - \beta_2)(z_2 - 1)P(0, 0, 0) \\ +[A(z_1, z_2)(1 - \beta_2 + \beta_2 z_2) - z_2]P(0, 0, z_2) \end{array} \right\}}{z_2 A(z_1, z_2)(z_1 - E_1(z_1, z_2))}, \tag{4.37}$$

where we have used equation (4.36) to eliminate $R(0, z_2)$.

Next, we also notice that the partial pgf $R(z_1, z_2)$ must be bound for all values of z_j such that $|z_j| < 1$ ($j = 1, 2$). In particular, this should be true for $z_1 = Y_1(z_2)$, with - as before -

$$Y_1(z) \triangleq E_1(Y_1(z), z), \tag{4.38}$$

and $|z_2| < 1$. If we put $z_1 = Y_1(z_2)$ in equation (4.37), the denominator of the right-hand side of this equation equals zero. The same must then be true for its numerator, yielding

$$P(0,0,z_2) = P(0,0,0)\frac{A(Y_1(z_2),z_2)(1-\beta_2)(z_2-1)}{z_2 - A(Y_1(z_2),z_2)(1-\beta_2+\beta_2 z_2)}. \tag{4.39}$$

Using this equation and equations (4.36)-(4.37) in equation (4.35), the following expression for $P(x_1,z_1,z_2)$ is derived:

$$\begin{aligned} P(x_1,z_1,z_2) = &\frac{P(0,0,0)(1-\beta_2)(z_2-1)}{z_2 - A(Y_1(z_2),z_2)(1-\beta_2+\beta_2 z_2)}\Bigg[A(Y_1(z_2),z_2) \\ &- \frac{x_1 z_1 (A(Y_1(z_2),z_2) - A(z_1,z_2))(S_1(x_1) - E_1(z_1,z_2))}{(x_1 - A(z_1,z_2))(z_1 - E_1(z_1,z_2))}\Bigg]. \end{aligned} \tag{4.40}$$

Finally, in order to find an expression for $P(0,0,0)$, we put $x_1 = z_1 = z_2 = 1$ and use de l'Hôpital's rule in equation (4.40). Using

$$Y_1'(1) = \frac{\lambda_2 \mu_1}{1-\rho_1}, \tag{4.41}$$

we find the expected result

$$P(0,0,0) = 1 - \rho_T. \tag{4.42}$$

If we assume geometric class-1 service times with parameter β_1, $P(1,z_1,z_2)$ equals expression (4.16) of $U(z_1,z_2)$ in the previous model, as expected.

4.2.3 Generally distributed service times

Finally, we analyze the system contents in case of generally distributed service times for both classes (which can be different for the two classes). The set $\{(r_{1,k},u_{1,k},u_{2,k}), k \geq 1\}$, as defined in the previous subsection, does no longer form a Markov chain, since the class-2 service times are also generally distributed in this subsection. Therefore, we introduce a new series of stochastic variables $r_{2,k}$ as follows: $r_{2,k}$ indicates the remaining number of slots service time of the oldest class-2 packet in the system at the beginning of slot k, if $u_{2,k} > 0$, and $r_{2,k} = 0$ if $u_{2,k} = 0$. A sample of the time axis is given in Figure 4.2 in order to demonstrate the relevant stochastic variables.

With this definition, $\{(r_{1,k},u_{1,k},r_{2,k},u_{2,k}), k \geq 1\}$ is seen to constitute a Markovian state description of the system at the beginning of slot k. Let $s^*_{j,k}$ ($j = 1,2$) indicate the service time of the next class-j packet to receive service at the beginning of slot k. The following system equations are established:

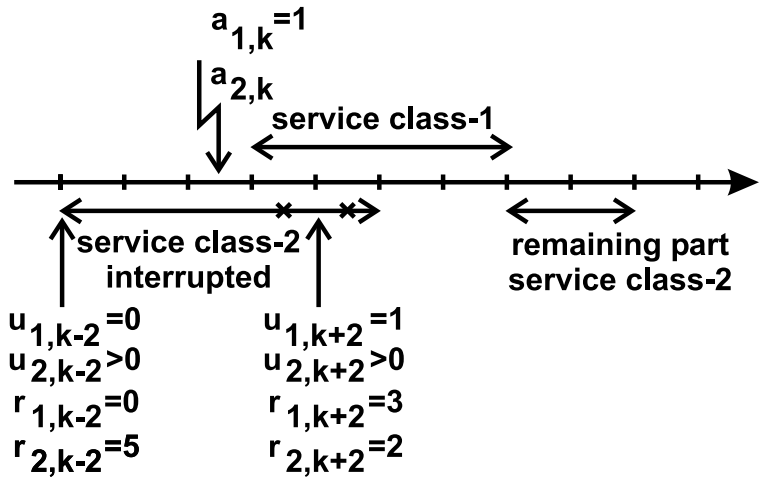


Figure 4.2: Sample of the time-axis for general class-1 and class-2 service times

1. If $r_{1,k} = 0$ (and hence $u_{1,k} = 0$):

   Class-1 system equations:

$$u_{1,k+1} = a_{1,k} \tag{4.43}$$

$$r_{1,k+1} = \begin{cases} 0 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases}. \tag{4.44}$$

   The only class-1 packets present in the system at the beginning of slot $k+1$ are the packets that arrive during the previous slot. If there have been new arrivals of class-1 packets during slot k, the remaining number of slots needed to serve the first class-1 packet is that packet's full service time.

   Class-2 system equations:

   In order to give expressions for $u_{2,k+1}$ and $r_{2,k+1}$, the value of $r_{2,k}$ is important. Three situations are distinguished:

   (a) If $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$u_{2,k+1} = a_{2,k} \tag{4.45}$$

$$r_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases}, \tag{4.46}$$

i.e., when class-2 packets arrive during slot k, the remaining service time of the oldest class-2 packet is this packet's service time.

(b) If $r_{2,k} = 1$:

$$u_{2,k+1} = u_{2,k} - 1 + a_{2,k}; \tag{4.47}$$

$$r_{2,k+1} = \begin{cases} 0 & \text{if } u_{2,k} - 1 + a_{2,k} = 0 \\ s^*_{2,k} & \text{if } u_{2,k} - 1 + a_{2,k} > 0 \end{cases}. \tag{4.48}$$

The class-2 packet in service at the beginning of slot k leaves the system at the end of slot k. If there are more class-2 packets present at the end of slot k, the remaining number of slots needed to serve the oldest class-2 packet is that packet's complete service time.

(c) If $r_{2,k} > 1$:

$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{4.49}$$

$$r_{2,k+1} = r_{2,k} - 1, \tag{4.50}$$

i.e., the class-2 packet in service at the beginning of slot k remains in the system (not necessarily in the server - it only remains in the server if there are no new class-1 arrivals). Its remaining service time is decreased by one.

2. If $r_{1,k} > 0$:

Class-1 system equations:

(a) If $r_{1,k} = 1$:

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \tag{4.51}$$

$$r_{1,k+1} = \begin{cases} 0 & \text{if } u_{1,k} - 1 + a_{1,k} = 0 \\ s^*_{1,k} & \text{if } u_{1,k} - 1 + a_{1,k} > 0 \end{cases}. \tag{4.52}$$

The class-1 packet in service at the beginning of slot k, leaves the system at the end of slot k. If any class-1 packets present, a new one enters the server.

(b) If $r_{1,k} > 1$:

$$u_{1,k+1} = u_{1,k} + a_{1,k} \tag{4.53}$$

$$r_{1,k+1} = r_{1,k} - 1, \tag{4.54}$$

i.e., the class-1 packet in service at the beginning of slot k stays in the server at the beginning of slot $k + 1$. Its remaining service time is decreased by one.

Class-2 system equations:

(a) If $r_{2,k} = 0$ (and hence $u_{2,k} = 0$):

$$u_{2,k+1} = a_{2,k} \tag{4.55}$$

$$r_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases}. \tag{4.56}$$

There were no class-2 packets in the system at the beginning of slot k.

(b) If $r_{2,k} > 0$:

$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{4.57}$$

$$r_{2,k+1} = r_{2,k}. \tag{4.58}$$

The difference with the previous situation is that there are class-2 packets in the system at the beginning of slot k. The remaining service time of the oldest class-2 packet stays the same.

We define $P_k(x_1, z_1, x_2, z_2)$ as the pgf of the state vector $(r_{1,k}, u_{1,k}, r_{2,k}, u_{2,k})$:

$$P_k(x_1, z_1, x_2, z_2) \triangleq \mathrm{E}[x_1^{r_{1,k}} z_1^{u_{1,k}} x_2^{r_{2,k}} z_2^{u_{2,k}}]. \tag{4.59}$$

Note that we use the same notation for the 4-dimensional pgf in this subsection, as we have used for the 3-dimensional pgf in the previous subsection. This is done because they are both the pgf of the Markovian state vectors. Since these are two different analyses, this does not give rise to any confusion.

Using the system equations (4.43)-(4.58), we constitute a relation between $P_k(x_1, z_1, x_2, z_2)$ and $P_{k+1}(x_1, z_1, x_2, z_2)$- taking into account the statistical independence of the random variables $(r_{1,k}, u_{1,k}, r_{2,k}, u_{2,k})$, $(a_{1,k}, a_{2,k})$, $s^*_{1,k}$ and $s^*_{2,k}$ respectively - which is given by

$$\begin{aligned} &P_{k+1}(x_1, z_1, x_2, z_2) \qquad (4.60)\\ &= \big[A(0,0) + (A(0,z_2) - A(0,0))S_2(x_2) + (A(z_1,0) - A(0,0))S_1(x_1) \\ &\quad + (A(z_1,z_2) - A(z_1,0) - A(0,z_2) + A(0,0))S_1(x_1)S_2(x_2)\big]P_k(0,0,0,0) \\ &\quad + A(0,0)R_{2,k}(0) + \big[A(0,z_2)R_{2,k}(z_2) - A(0,0)R_{2,k}(0)\big]S_2(x_2) \\ &\quad + (A(z_1,0) - A(0,0))R_{2,k}(0)S_1(x_1) + \big[(A(z_1,z_2) - A(0,z_2))R_{2,k}(z_2) \\ &\quad - (A(z_1,0) - A(0,0))R_{2,k}(0)\big]S_1(x_1)S_2(x_2) \\ &\quad + \frac{A(0,z_2) + (A(z_1,z_2) - A(0,z_2))S_1(x_1)}{x_2}\big[P_k(0,0,x_2,z_2) - x_2 z_2 R_{2,k}(z_2) \\ &\quad - P_k(0,0,0,0)\big] + \big[A(0,0) + (A(0,z_2) - A(0,0))S_2(x_2)\big]R_{1,k}(0,0,0) \\ &\quad + \big[A(z_1,0)R_{1,k}(z_1,0,0) - A(0,0)R_{1,k}(0,0,0)\big]S_1(x_1) \end{aligned}$$

$$
\begin{aligned}
&+\big[(A(z_1,z_2)-A(z_1,0))R_{1,k}(z_1,0,0)-(A(0,z_2)-A(0,0))R_{1,k}(0,0,0)\big]\\
&\times S_1(x_1)S_2(x_2)+A(0,z_2)\big[R_{1,k}(0,x_2,z_2)-R_{1,k}(0,0,0)\big]\\
&+\big[A(z_1,z_2)(R_{1,k}(z_1,x_2,z_2)-R_{1,k}(z_1,0,0))\\
&-A(0,z_2)(R_{1,k}(0,x_2,z_2)-R_{1,k}(0,0,0))\big]S_1(x_1)\\
&+\frac{A(z_1,0)+(A(z_1,z_2)-A(z_1,0)S_2(x_2)}{x_1}\\
&\times\big[P_k(x_1,z_1,0,0)-x_1z_1R_{1,k}(z_1,0,0)-P_k(0,0,0,0)\big]\\
&+\frac{A(z_1,z_2)}{x_1}\big[(P_k(x_1,z_1,x_2,z_2)-P_k(x_1,z_1,0,0))-x_1z_1(R_{1,k}(z_1,x_2,z_2)\\
&-R_{1,k}(z_1,0,0))-(P_k(0,0,x_2,z_2)-P_k(0,0,0,0))\big],
\end{aligned}
$$

with the partial pgf's $R_{1,k}(z_1,x_2,z_2)$ and $R_{2,k}(z_2)$ defined as

$$R_{1,k}(z_1,x_2,z_2)\triangleq\mathrm{E}\left[z_1^{u_{1,k}-1}x_2^{r_{2,k}}z_2^{u_{2,k}}\{r_{1,k}=1\}\right] \tag{4.61}$$

$$R_{2,k}(z_2)\triangleq\mathrm{E}\left[z_2^{u_{2,k}-1}\{r_{1,k}=u_{1,k}=0,r_{2,k}=1\}\right]. \tag{4.62}$$

We assume that the queueing system is stable and as a result $P_k(x_1,z_1,x_2,z_2)$, $R_{1,k}(z_1,x_2,z_2)$ and $R_{2,k}(z_2)$ converge to the steady-state values:

$$P(x_1,z_1,x_2,z_2)\triangleq\lim_{k\to\infty}P_k(x_1,z_1,x_2,z_2) \tag{4.63}$$

$$R_1(z_1,x_2,z_2)\triangleq\lim_{k\to\infty}R_{1,k}(z_1,x_2,z_2) \tag{4.64}$$

$$R_2(z_2)\triangleq\lim_{k\to\infty}R_{2,k}(z_2), \tag{4.65}$$

respectively. By taking the $k\to\infty$ limit in (4.60) we obtain

$$
\begin{aligned}
&P(x_1,z_1,x_2,z_2) \qquad (4.66)\\
&=\frac{1}{x_2(x_1-A(z_1,z_2))}\Big\{\big[x_1x_2A(0,0)(1-S_1(x_1))(1-S_2(x_2))\\
&+x_1A(0,z_2)(1-S_1(x_1))(x_2S_2(x_2)-1)\\
&+x_2A(z_1,0)(x_1S_1(x_1)-1)(1-S_2(x_2))\\
&+A(z_1,z_2)(x_1S_1(x_1)(x_2S_2(x_2)-1)-x_2(S_2(x_2)-1))\big]P(0,0,0,0)\\
&+x_1x_2\big[A(0,0)(1-S_1(x_1))+A(z_1,0)S_1(x_1)\big](1-S_2(x_2))R_2(0)\\
&+x_1x_2(A(0,z_2)-A(0,0))(1-S_1(x_1))(S_2(x_2)-1)R_1(0,0,0)\\
&+x_2(A(z_1,z_2)-A(z_1,0))(S_2(x_2)-1)P(x_1,z_1,0,0)\\
&+x_1x_2(A(z_1,z_2)-A(z_1,0))(z_1-S_1(x_1))(1-S_2(x_2))R_1(z_1,0,0)\\
&+\big[x_1A(0,z_2)(1-S_1(x_1))+A(z_1,z_2)(x_1S_1(x_1)-x_2)\big]P(0,0,x_2,z_2)\\
&+x_1x_2\big[A(0,z_2)(1-S_1(x_1))+A(z_1,z_2)S_1(x_1)\big](S_2(x_2)-z_2)R_2(z_2)
\end{aligned}
$$

$$+ x_1x_2A(0,z_2)(1-S_1(x_1))R_1(0,x_2,z_2)$$
$$+ x_1x_2A(z_1,z_2)(S_1(x_1)-z_1)R_1(z_1,x_2,z_2)\Big\}.$$

It now remains for us to determine the unknown functions $P(x_1,z_1,0,0)$, $R_1(z_1,0,0)$, $P(0,0,x_2,z_2)$, $R_2(z_2)$, $R_1(0,x_2,z_2)$ and $R_1(z_1,x_2,z_2)$ and the unknown parameters $P(0,0,0,0)$, $R_2(0)$ and $R_1(0,0,0)$. This is done in a few steps.

We observe that, due to the fact that $r_{j,k}=0$ iff $u_{j,k}=0$ ($j=1,2$ respectively), the following equations hold:

$$P(x_1,0,x_2,0) = P(0,0,0,0) \tag{4.67}$$
$$R_1(0,x_2,0) = R_1(0,0,0), \tag{4.68}$$

for all x_j ($j=1,2$). By putting $z_j=0$ ($j=1,2$) in (4.66) and using equations (4.67) and (4.68), we obtain the following relation between $P(0,0,0,0)$, $R_2(0)$ and $R_1(0,0,0)$:

$$P(0,0,0,0) = A(0,0)[P(0,0,0,0)+R_2(0)+R_1(0,0,0)]. \tag{4.69}$$

We furthermore observe that the following equations hold - again because $r_{j,k}=0$ iff $u_{j,k}=0$ -

$$P(x_1,z_1,x_2,0) = P(x_1,z_1,0,0) \tag{4.70}$$
$$R_1(z_1,x_2,0) = R_1(z_1,0,0) \tag{4.71}$$
$$P(x_1,0,x_2,z_2) = P(0,0,x_2,z_2), \tag{4.72}$$

for all x_j and z_j ($j=1,2$). Replacing z_2 (z_1 respectively) by 0 in equation (4.66) and using the former equations and equation (4.69), we find the following expressions for $P(x_1,z_1,0,0)$ and $P(0,0,x_2,z_2)$ respectively:

$$P(x_1,z_1,0,0) = \frac{\left\{\begin{array}{l}\big[x_1(1-S_1(x_1))+A(z_1,0)(x_1S_1(x_1)-1)\big]P(0,0,0,0)\\ +x_1A(z_1,0)\big[S_1(x_1)R_2(0)+(S_1(x_1)-z_1)R_1(z_1,0,0)\big]\end{array}\right\}}{x_1-A(z_1,0)} \tag{4.73}$$

$$P(0,0,x_2,z_2) = \frac{\left\{\begin{array}{l}\big[x_2(1-S_2(x_2))+A(0,z_2)(x_2S_2(x_2)-1)\big]P(0,0,0,0)\\ +x_2A(0,z_2)(S_2(x_2)-1)R_1(0,0,0)\\ +x_2A(0,z_2)\big[(S_2(x_2)-z_2)R_2(z_2)+R_1(0,x_2,z_2)\big]\end{array}\right\}}{x_2-A(0,z_2)}. \tag{4.74}$$

Substituting these two expressions in expression (4.66) allows us to eliminate $P(x_1,z_1,0,0)$ and $P(0,0,x_2,z_2)$ in this latter expression.

Firstly, we determine all unknown functions in the right-hand side of expression (4.74) as a function of the - at this moment still unknown - constant $P(0,0,0,0)$. Therefore, we go back to expression (4.66). We notice that the function $P(x_1,z_1,x_2,z_2)$ must be bound for all values of x_j and z_j such that $|x_j|<1$ and $|z_j|<1$ ($j=1,2$) since $P(x_1,z_1,x_2,z_2)$ is a pgf. In particular, this should be true for $x_1=A(z_1,z_2)$, $|x_2|<1$ and $|z_j|<1$ ($j=1,2$), since $|A(z_1,z_2)|<1$ for all such x_2 and z_j. The above implies that if we choose $x_1=A(z_1,z_2)$ in equation (4.66), where $|x_2|<1$ and $|z_j|<1$, the denominator of the right-hand side of this equation vanishes. Of course, the same must then be true for the numerator, which yields the following expression:

$$\begin{aligned}(S_2(x_2)-1)&R_1(z_1,0,0)+R_1(z_1,x_2,z_2) \qquad (4.75)\\ =&\frac{1}{A(z_1,z_2)(x_2-A(0,z_2))(z_1-E_1(z_1,z_2))}\\ &\times\Big\{(A(z_1,z_2)-A(0,z_2))E_1(z_1,z_2)S_2(x_2)(x_2-1)P(0,0,0,0)\\ &+A(0,z_2)(A(z_1,z_2)-x_2)E_1(z_1,z_2)(S_2(x_2)-1)R_1(0,0,0)\\ &+x_2(A(z_1,z_2)-A(0,z_2))E_1(z_1,z_2)(S_2(x_2)-z_2)R_2(z_2)\\ &+A(0,z_2)(A(z_1,z_2)-x_2)E_1(z_1,z_2)R_1(0,x_2,z_2)\Big\}.\end{aligned}$$

Here we also substituted $P(x_1,z_1,0,0)$ and $P(0,0,x_2,z_2)$ by their expressions obtained in equations (4.73) and (4.74) respectively.

Next, we notice that $(S_2(x_2)-1)R_1(z_1,0,0)+R_1(z_1,x_2,z_2)$ must be bound for all values of x_2 and z_j such that $|x_2|<1$ and $|z_j|<1$ ($j=1,2$). In particular, this should be true for $z_1=Y_1(z_2)$, with

$$Y_1(z)\triangleq E_1(Y_1(z),z). \qquad (4.76)$$

The above implies that if we insert $z_1=Y_1(z_2)$ in equation (4.75), where $|z_2|<1$, the denominator of the right-hand side of this equation vanishes. The same must then be true for its numerator, yielding

$$\begin{aligned}(S_2(x_2)-1)&R_1(0,0,0)+R_1(0,x_2,z_2) \qquad (4.77)\\ =&\frac{A(Y_1(z_2),z_2)-A(0,z_2)\left\{\begin{matrix}S_2(x_2)(x_2-1)P(0,0,0,0)\\ +x_2(S_2(x_2)-z_2)R_2(z_2)\end{matrix}\right\}}{A(0,z_2)(x_2-A(Y_1(z_2),z_2))}.\end{aligned}$$

Next, we notice that $(S_2(x_2)-1)R_1(0,0,0)+R_1(0,x_2,z_2)$ must be bounded for all values of x_2 and z_2 such that $|x_2|<1$ and $|z_2|<1$. In particular, this should be true for $x_2=A(Y_1(z_2),z_2)$. The above implies that if we choose $x_2=A(Y_1(z_2),z_2)$ in equation (4.77), the denominator of the right-hand side of this equation vanishes. The same must then be true for its numerator, yielding

the following expression for $R_2(z_2)$:

$$R_2(z_2) = \frac{P(0,0,0,0)Y_2(z_2)(A(Y_1(z_2),z_2)-1)}{A(Y_1(z_2),z_2)(z_2-Y_2(z_2))}, \tag{4.78}$$

with

$$Y_2(z) \triangleq S_2(A(Y_2(z),z)). \tag{4.79}$$

Using equations (4.77) and (4.78) in equation (4.74), we find:

$$\begin{aligned} P(0,0,x_2,z_2) =& P(0,0,0,0) \\ &\times \left[1 + x_2 z_2 \frac{(A(Y_1(z_2),z_2)-1)(S_2(x_2)-Y_2(z_2))}{(x_2 - A(Y_1(z_2),z_2))(z_2-Y_2(z_2))}\right]. \end{aligned} \tag{4.80}$$

It now remains for us to determine the unknown functions $P(x_1,z_1,0,0)$, $R_1(z_1,0,0)$ and the unknown parameters $P(0,0,0,0)$, $R_2(0)$ and $R_1(0,0,0)$ - we already have relation (4.69) between the latter three constants - in expression (4.66). Therefore, we return to expression (4.73). We notice that the function $P(x_1,z_1,0,0)$ must be bound for all values of x_1 and z_1 such that $|x_1|<1$ and $|z_1|<1$ since $P(x_1,z_1,x_2,z_2)$ is a pgf. In particular, this should be true for $x_1 = A(z_1,0)$, $|z_1|<1$, since $|A(z_1,0)|<1$ for all such z_1, because $A(z_1,z_2)$ is a pgf. The above implies that if we choose $x_1 = A(z_1,0)$ in equation (4.73), where $|z_1|<1$, the denominator of the right-hand side of this equation vanishes. Of course, the same must then be true for its numerator, which yields the following relation for $R_1(z_1,0,0)$:

$$R_1(z_1,0,0) = \frac{E_1(z_1,0)[(A(z_1,0)-1)P(0,0,0,0) + A(z_1,0)R_2(0)]}{A(z_1,0)(z_1 - E_1(z_1,0))}. \tag{4.81}$$

Finally, $R_2(0)$ is easily calculated by substituting z_2 by 0 in expression (4.78)

$$R_2(0) = P(0,0,0,0)\frac{1-A(Y_1(0),0)}{A(Y_1(0),0)}. \tag{4.82}$$

We find the following expression for $P(x_1,z_1,0,0)$ from equation (4.73) together with equations (4.81) and (4.82):

$$\begin{aligned} P(x_1,z_1,0,0) =& P(0,0,0,0) \\ &\times \left[1 + x_1 z_1 \frac{(A(z_1,0)-A(Y_1(0),0))(S_1(x_1)-S_1(A(z_1),0))}{A(Y_1(0),0)(x_1 - A(z_1,0))(z_1-E_1(z_1,0))}\right]. \end{aligned} \tag{4.83}$$

The following expression for $P(x_1,z_1,x_2,z_2)$ as function of the system parameters, the functions $Y_j(z)$ and $P(0,0,0,0)$ is now derived by substituting all

found unknown functions and constants in expression (4.66):

$$\begin{aligned} P(x_1,z_1,x_2,z_2) =& P(0,0,0,0)\Bigg[1+ \\ & \frac{x_1 z_1 (A(z_1,0)-A(Y_1(0),0))(S_1(x_1)-E_1(z_1,0))(1-S_2(x_2))}{A(Y_1(0),0)(x_1-A(z_1,0))(z_1-E_1(z_1,0))} \\ & + x_1 z_1 \frac{(A(z_1,z_2)-A(Y_1(z_2),z_2))(S_1(x_1)-E_1(z_1,z_2))}{(x_1-A(z_1,z_2))(z_1-E_1(z_1,z_2))(z_2-Y_2(z_2))} \\ & \times \left\{ \frac{Y_2(z_2)(z_2-S_2(x_2))}{A(Y_1(z_2),z_2)} + z_2 \frac{(x_2-1)(S_2(x_2)-Y_2(z_2))}{x_2-A(Y_1(z_2),z_2)} \right\} \\ & + x_2 z_2 \frac{(A(Y_1(z_2),z_2)-1)(S_2(x_2)-Y_2(z_2))}{(x_2-A(Y_1(z_2),z_2))(z_2-Y_2(z_2))} \Bigg]. \end{aligned} \tag{4.84}$$

Finally, in order to find an expression for $P(0,0,0,0)$, we put $x_1 = z_1 = x_2 = z_2 = 1$ and use de l'Hôpital's rule in equation (4.84). Again the value of $Y_1'(1)$

$$Y_1'(1) = \frac{\lambda_2 \mu_1}{1-\rho_1}, \tag{4.85}$$

is needed in this calculation. We obtain the expected result for the probability of an empty system:

$$P(0,0,0,0) = 1-\rho_T. \tag{4.86}$$

Substituting this result in equation (4.84), we finally obtain a fully determined - albeit an elaborate - expression for $P(x_1,z_2,x_2,z_2)$.

If we assume geometric class-2 service times, $P(x_1,z_1,1,z_2)$ equals expression (4.40), as expected.

This concludes the calculation of the multivariate pgf's for the three different models, which are the starting-points for all further calculations. However, since these calculations are fairly similar for these three models, we will in the remainder of the analysis only concentrate on the most extensive model discussed in this subsection. From this point on, the service times of both classes are thus assumed to be generally distributed and we start from expression (4.84) of the multivariate pgf $P(x_1,z_1,x_2,z_2)$. The calculation of some of the performance measures for the other two models can be found in [Walraevens et al. 2000a, 2004b] and [Walraevens et al. 2001, 2004a] respectively.

4.3 System contents

In this section, we calculate marginal pgf's, moments and tail probabilities of the steady-state total, class-1 and class-2 system contents.

4.3.1 Calculation of the pgf $P_1(x,z)$

From expression (4.84) some useful joint pgf's are calculated. First, we calculate the joint steady-state pgf of the system contents of class-1 packets and the residual service time of the class-1 packet in service:

$$P_1(x,z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[x^{r_{1,k}} z^{u_{1,k}}\right] \tag{4.87}$$

$$= P(x,z,1,1) \tag{4.88}$$

$$= (1-\rho_1)\left[1 + xz\frac{(A_1(z)-1)(S_1(x)-S_1(A_1(z)))}{(x-A_1(z))(z-S_1(A_1(z)))}\right]. \tag{4.89}$$

This joint pgf is independent of class-2 characteristics, due to the preemptive property of the priority scheduling discipline. From the point-of-view of class-1 packets, it is as if they are the only packets in the system. This pgf is thus the same as obtained in Bruneel [1993], wherein a single-class system with the numbers of per-slot arrivals i.i.d. and with general service times is analyzed.

4.3.2 Calculation of the pgf $P_2(x,z)$

Secondly, we calculate the joint pgf of the system contents of class-2 packets and the remaining service time of the oldest class-2 packet at the beginning of a random slot in steady-state (note that this packet is not necessarily in service) from equation (4.84), yielding

$$P_2(x,z) \triangleq \lim_{k\to\infty} \mathrm{E}\left[x^{r_{2,k}} z^{u_{2,k}}\right] \tag{4.90}$$

$$= P(1,1,x,z) \tag{4.91}$$

$$\begin{aligned} = (1-\rho_T)\Bigg[&\frac{A_2(0)(1-A(Y_1(0),0)) - (A_2(0)-A(Y_1(0),0))S_2(x)}{A(Y_1(0),0)(1-A_2(0))} \\ &+ \frac{Y_2(z)(A_2(z)-A(Y_1(z),z))(z-S_2(x))}{A(Y_1(z),z)(1-A_2(z))(z-Y_2(z))} \\ &+ z\frac{(x-1)(A_2(z)-A(Y_1(z),z))(S_2(x)-Y_2(z))}{(1-A_2(z))(x-A(Y_1(z),z))(z-Y_2(z))} \\ &+ xz\frac{(A(Y_1(z),z)-1)(S_2(x)-S_2(A(Y_1(z),z)))}{(x-A(Y_1(z),z))(z-Y_2(z))}\Bigg]. \end{aligned} \tag{4.92}$$

4.3.3 Calculation of the pgf $U(z_1, z_2)$

Thirdly - and most importantly - we calculate the joint pgf of the system contents of class-1 and class-2 packets at the beginning of a random slot from equation (4.84). It is given by:

$$\begin{aligned} U(z_1, z_2) \triangleq & \lim_{k\to\infty} \mathrm{E}\left[z_1^{u_{1,k}} z_2^{u_{2,k}}\right] && (4.93)\\ =& P(1, z_1, 1, z_2) && (4.94)\\ =& (1-\rho_T)\frac{Y_2(z_2)(z_2-1)}{z_2 - Y_2(z_2)} && (4.95)\\ & \times \left[1 + z_1 \frac{(A(z_1, z_2) - A(Y_1(z_2), z_2))(E_1(z_1, z_2) - 1)}{A(Y_1(z_2), z_2)(A(z_1, z_2) - 1)(z_1 - E_1(z_1, z_2))}\right]. \end{aligned}$$

4.3.4 The marginal pgf $U_T(z)$

From expression (4.95) of the two-dimensional pgf $U(z_1, z_2)$, we derive an expression for the pgf of the total steady-state system contents at the beginning of an arbitrary slot, yielding

$$\begin{aligned} U_T(z) \triangleq & \lim_{k\to\infty} \mathrm{E}\left[z^{u_{T,k}}\right] && (4.96)\\ =& U(z, z) && (4.97)\\ =& (1-\rho_T)\frac{Y_2(z)(z-1)}{z - Y_2(z)} && (4.98)\\ & \times \left[1 + z \frac{(A_T(z) - A(Y_1(z), z))(S_1(A_T(z)) - 1)}{A(Y_1(z), z)(A_T(z) - 1)(z - S_1(A_T(z)))}\right]. \end{aligned}$$

Note that in the special case that the service times of both classes are equally distributed, this expression does no longer equal the pgf of the system contents in a single-class queue, as is the case for a queue with an *NP* priority scheduling discipline. The reason is that, since the class-2 service times can be interrupted, 'partial' packets are stored in the queue, which are counted as (integer) packets in the system contents.

4.3.5 The marginal pgf $U_1(z)$

The pgf of the class-1 system contents is calculated from $U(z_1, z_2)$ as follows:

$$\begin{aligned} U_1(z) \triangleq & \lim_{k\to\infty} \mathrm{E}\left[z^{u_{1,k}}\right] && (4.99)\\ =& U(z, 1) && (4.100)\\ =& (1-\rho_1)\frac{S_1(A_1(z))(z-1)}{z - S_1(A_1(z))}. && (4.101) \end{aligned}$$

Note that $U_1(z)$ is also obtained from $P_1(x,z)$ by substituting x by 1.

This expression is identical to the pgf of the system contents in a single-class buffer with $A_1(z)$ the pgf of the number of per-slot arrivals and the pgf of the service times equal to $S_1(z)$. Indeed - as already discussed - since the priority scheduling discipline is *preemptive*, the class-1 system contents is not influenced by arriving class-2 traffic.

4.3.6 The marginal pgf $U_2(z)$

The pgf of the class-2 system contents is easily determined from $U(z_1,z_2)$:

$$\begin{aligned} U_2(z) \triangleq& \lim_{k\to\infty} \mathrm{E}\left[z^{u_{2,k}}\right] && (4.102)\\ =&U(1,z) && (4.103)\\ =&(1-\rho_T)\frac{Y_2(z)(z-1)}{z-Y_2(z)}\frac{A_2(z)(A(Y_1(z),z)-1)}{A(Y_1(z),z)(A_2(z)-1)}. && (4.104) \end{aligned}$$

Alternatively, this expression is found by substituting x by 1 in $P_2(x,z)$.

4.3.7 Calculation of moments

From the pgf's (4.98), (4.101) and (4.104), we calculate the moments of the total, class-1 and class-2 system contents respectively. We give the expressions of the means in this subsection.

The mean total system contents is given by

$$\begin{aligned} \mathrm{E}[u_T] =&U_T'(1) && (4.105)\\ =&\frac{\rho_T}{2}+\frac{\mu_1\mathrm{Var}[a_T]}{2(1-\rho_T)}-\frac{\mu_1\lambda_2(\mu_2-\mu_1)\mathrm{Var}[a_1]}{2(1-\rho_T)(1-\rho_1)}+\frac{(\mu_2-\mu_1)\mathrm{Var}[a_2]}{2(1-\rho_T)} && (4.106)\\ &+\frac{\lambda_1\mathrm{Var}[s_1](\lambda_1(1-\rho_1)+\lambda_2(1-\mu_2\lambda_1))}{2(1-\rho_T)(1-\rho_1)}+\frac{\lambda_2^2\mathrm{Var}[s_2]}{2(1-\rho_T)(1-\rho_1)}\\ &+\frac{\rho_1\lambda_2(\mu_2-1)}{2(1-\rho_1)}. \end{aligned}$$

The mean class-1 system contents is found by taking the first derivative of $U_1(z)$ and substituting z by 1, yielding

$$\begin{aligned} \mathrm{E}[u_1] =&U_1'(1) && (4.107)\\ =&\frac{\rho_1}{2}+\frac{\mu_1\mathrm{Var}[a_1]}{2(1-\rho_1)}+\frac{\lambda_1^2\mathrm{Var}[s_1]}{2(1-\rho_1)}. && (4.108) \end{aligned}$$

Finally, the mean class-2 system contents is given by

$$\begin{aligned}\mathrm{E}[u_2] =& U_2'(1) && (4.109)\\ =& \frac{\rho_2}{2} + \frac{\mu_1^2\lambda_2\mathrm{Var}[a_1]}{2(1-\rho_T)(1-\rho_1)} + \frac{\mu_2\mathrm{Var}[a_2]}{2(1-\rho_T)} + \frac{\mu_1\mathrm{Cov}[a_1,a_2]}{1-\rho_T} && (4.110)\\ &+ \frac{\lambda_2(\lambda_1\mathrm{Var}[s_1]+\lambda_2\mathrm{Var}[s_2])}{2(1-\rho_T)(1-\rho_1)} + \frac{\rho_1\lambda_2(\mu_2-1)}{2(1-\rho_1)}.\end{aligned}$$

Note that expressions (4.106), (4.108) and (4.110) satisfy $\mathrm{E}[u_T] = \mathrm{E}[u_1] + \mathrm{E}[u_2]$, since $\mathrm{Var}[a_T]$ =Var$[a_1]$+Var$[a_1]$ + 2Cov$[a_1, a_2]$.

4.3.8 Calculation of tail probabilities

As in the previous chapters, approximate tail probabilities of the system contents can be calculated, i.e., approximate expressions of Prob$[u = n]$ and Prob$[u > L]$ (with u the class-1 or class-2 system contents respectively) are found for large enough n and L.

The respective pgf's of the system contents show a similar qualitative behavior as the pgf's of the system contents in the case of an NP priority scheduling. More precisely, exactly the same singularities play a role in the pgf's for both types of priority. The difference in the behavior of the pgf's near their dominant singularity (and thus in the tail probabilities) are the scaling factors K_T, K_1 and the $K_2^{(i)}$'s ($i = 1, 2, 3$) (see further). We will thus only show the expressions of the tail probabilities in this subsection and refer to subsections 2.1.7 and 3.3.6 for (more) detailed derivations. We once again note however that we assume that the pgf's of the arrival and service processes ($A_T(z)$, the $A_j(z)$ and the $S_j(z)$) and their derivatives go to infinity for z equal to their radii of convergence or for $z \to \infty$.

Behavior of $Y_1(z)$ and $Y_2(z)$

$Y_1(z)$ is implicitly defined as $Y_1(z) = S_1(A(Y_1(z), z))$. This function has a dominant branch-point singularity z_B where $Y'(z)$ becomes infinite, i.e.,

$$E_1^{(1)}(Y_1(z_B), z_B) = 1, \tag{4.111}$$

with $E_1(z_1, z_2) \triangleq S_1(A(z_1, z_2))$ and $E_1^{(j)}(x, y) = \left.\dfrac{\partial E_1(z_1, z_2)}{\partial z_j}\right|_{z_1=x, z_2=y}$, $j = 1, 2$. In the neighborhood of this (dominant) singularity, $Y_1(z)$ is approximately given by

$$Y_1(z) \approx Y_1(z_B) - K_{Y_1}\left(z_B - z\right)^{1/2}, \tag{4.112}$$

with

$$K_{Y_1} = \sqrt{\frac{2A^{(2)}(Y_1(z_B), z_B)}{A^{(11)}(Y_1(z_B), z_B) + S_1''(A(Y_1(z_B), z_B)) \left(A^{(1)}(Y_1(z_B), z_B)\right)^3}}. \quad (4.113)$$

z_B is also a branch-point of $Y_2(z)$, since

$$Y_2'(z) = S_2'(A(Y_1(z), z)) \left[A^{(1)}(Y_1(z), z)Y_1'(z) + A^{(2)}(Y_1(z), z)\right]. \quad (4.114)$$

Thus in the neighborhood of z_B, we find

$$Y_2(z) \approx Y_2(z_B) - K_{Y_2}\sqrt{z_B - z}, \quad (4.115)$$

with

$$K_{Y_2} = K_{Y_1} S_2'(A(Y_1(z_B), z_B))A^{(1)}(Y_1(z_B), z_B). \quad (4.116)$$

Class-1 system contents

The dominant singularity z_H of $U_1(z)$ is the dominant zero of $z - S_1(A_1(z))$ on the positive real axis (> 1) and this singularity is a single pole. We get

$$u_1(n) \triangleq \text{Prob}[u_1 = n] \quad (4.117)$$
$$\approx K_1 z_H^{-n-1}, \quad (4.118)$$

for large enough n and

$$\text{Prob}[u_1 > L] \approx \frac{K_1 z_H^{-L-1}}{z_H - 1}, \quad (4.119)$$

for large enough L. The constant K_1 is given by

$$K_1 = \frac{(1 - \rho_1)(z_H - 1)z_H}{S_1'(A_1(z_H))A_1'(z_H) - 1}. \quad (4.120)$$

Class-2 system contents

$U_2(z)$ has 2 important singularities, a single pole z_L and a branch point z_B, with z_L the dominant zero of $z - Y_2(z)$ and z_B the branch-point of $Y_1(z)$ (thus $E_1^{(1)}(Y_1(z_B), z_B) = 1$). The tail behavior of the system contents of class-2 packets is thus characterized by z_L or z_B, depending on which is the dominant (i.e.,

smallest) singularity. The following three different types of tail behavior may occur:

$$u_2(n) \triangleq \text{Prob}[u_2 = n] \tag{4.121}$$

$$\approx \begin{cases} K_2^{(1)} z_L^{-n-1} & \text{if } z_L \text{ dominant} \\ \dfrac{K_2^{(2)} n^{-1/2} z_B^{-n}}{\sqrt{z_B \pi}} & \text{if } z_L = z_B \text{ dominant} \\ \dfrac{K_2^{(3)}}{2} \sqrt{\dfrac{z_B}{\pi}} n^{-3/2} z_B^{-n} & \text{if } z_B \text{ dominant,} \end{cases} \tag{4.122}$$

The constants $K_2^{(i)}$ are given by

$$K_2^{(1)} = \frac{(1-\rho_T) z_L A_2(z_L)(z_L - 1)(A(Y_1(z_L), z_L) - 1)}{A(Y_1(z_L), z_L)(A_2(z_L) - 1)\,(Y_2'(z_L) - 1)} \tag{4.123}$$

$$K_2^{(2)} = \frac{(1-\rho_T) z_B A_2(z_B)(z_B - 1)(A(Y_1(z_B), z_B) - 1)}{K_{Y_2} A(Y_1(z_B), z_B)(A_2(z_B) - 1)} \tag{4.124}$$

$$K_2^{(3)} = \frac{(1-\rho_T) A_2(z_B)(z_B - 1)}{A(Y_1(z_B), z_B)(A_2(z_B) - 1)(z_B - Y_2(z_B))} \times \left\{ \frac{K_{Y_1} A^{(1)}(Y_1(z_B), z_B) Y_2(z_B)}{A(Y_1(z_B), z_B)} + \frac{K_{Y_2} z_B (A(Y_1(z_B), z_B) - 1)}{z_B - Y_2(z_B))} \right\}. \tag{4.125}$$

Finally,

$$\text{Prob}[u_2 > L] \approx \frac{u_2(L)}{z_* - 1}, \tag{4.126}$$

with z_* the dominant singularity of $U_2(z)$.

4.4 Queue contents

The *queue contents* - defined as the number of packets in the queue (thus without the one in the server) - are easily derived from the system contents. We denote - as before - the queue contents of class-j at the beginning of the k-th slot by $q_{j,k}$ ($j = 1, 2$). The following relations between $q_{j,k}$ and $u_{j,k}$ are then found:

$$q_{1,k} = [u_{1,k} - 1]^+ \tag{4.127}$$

$$q_{2,k} = \begin{cases} [u_{2,k} - 1]^+ & \text{if } u_{1,k} = 0 \\ u_{2,k} & \text{if } u_{1,k} > 0 \end{cases}. \tag{4.128}$$

When a class-1 packet is in service during slot k - i.e. when $u_{1,k} > 0$, the class-1 queue contents is one less as the class-1 system contents, while the class-2 queue contents and class-2 system contents are equal. When a class-2 packet is in service during slot k ($u_{1,k} = 0, u_{2,k} > 0$), the class-1 system contents and class-1 queue contents are zero and the class-2 queue contents is one less than the class-2 system contents. This leads to the former equations. Taking the (two-dimensional) z-transform of these equations and letting $k \to \infty$ leads to

$$Q(z_1, z_2) = \lim_{k\to\infty} \mathrm{E}[z_1^{q_{1,k}} z_2^{q_{2,k}}] \tag{4.129}$$

$$= U(0,0) + \frac{U(0,z_2) - U(0,0)}{z_2} + \frac{U(z_1,z_2) - U(0,z_2)}{z_1}. \tag{4.130}$$

Substituting equation (4.95) in this expression yields

$$Q(z_1, z_2) = (1-\rho_T)\frac{z_2 - 1}{z_2 - Y_2(z_2)} \tag{4.131}$$

$$\times \left[1 + \frac{Y_2(z_2)(A(z_1,z_2) - A(Y_1(z_2),z_2))(E_1(z_1,z_2) - 1)}{A(Y_1(z_2),z_2)(A(z_1,z_2) - 1)(z_1 - E_1(z_1,z_2))}\right].$$

4.5 Unfinished work

The joint pgf of the steady-state unfinished work of class-1 and class-2 at the beginning of a random slot is easy to derive from the results in subsection 4.2.3. We denote the class-j unfinished work at the beginning of slot k by $w_{j,k}$, $j = 1, 2$. The following relations between $w_{j,k}$, $u_{j,k}$ and $r_{j,k}$ are established

$$w_{j,k} = r_{j,k} + \sum_{m=1}^{[u_{j,k}-1]^+} s_{j,m}, \tag{4.132}$$

$j = 1, 2$, with $r_{j,k}$ the residual service time of the oldest class-j packet at the beginning of slot k (as defined in subsection 4.2.3) and the $s_{j,m}$ ($m = 1, \ldots, [u_{j,k} - 1]^+$) the service times of the class-j packets in the system at the beginning of slot k (except for the oldest one). Expression (4.132) is explained as follows: all class-j packets except for the oldest one add a complete service time to the unfinished work, while the oldest class-j packet adds a residual service time.

By z-transforming equations (4.132), the joint pgf of the class-1 and class-2 steady-state unfinished work at the beginning of a random slot is calculated as a function of the four-dimensional pgf $P(x_1, z_1, x_2, z_2)$ of the class-1 residual service time and system contents and the class-2 residual service time and system contents respectively. We find

$$W(z_1, z_2) \triangleq \mathrm{E}[z_1^{w_{1,k}} z_2^{w_{2,k}}] \tag{4.133}$$

$$
\begin{aligned}
=&P(0,0,0,0) + \frac{P(z_1, S_1(z_1), 0, 0) - P(0,0,0,0)}{S_1(z_1)} \\
&+ \frac{P(0,0,z_2, S_2(z_2)) - P(0,0,0,0)}{S_2(z_2)} \\
&+ \frac{\left\{ \begin{array}{l} P(z_1, S_1(z_1), z_2, S_2(z_2)) - P(z_1, S_1(z_1), 0, 0) \\ -P(0,0,z_2, S_2(z_2)) + P(0,0,0,0) \end{array} \right\}}{S_1(z_1) S_2(z_2)}.
\end{aligned} \tag{4.134}
$$

The first term gives the partial pgf of the unfinished work when the system is empty, the second (third respectively) term is the partial pgf of the unfinished work of both classes when a class-1 (class-2 respectively) packet is being served while the class-2 (class-1 respectively) system contents are zero. Finally, the last term is the partial pgf of the unfinished work of both classes when there is at least one class-1 and at least one class-2 packet in the system.

Substituting expression (4.84) in expression (4.134) yields

$$
W(z_1, z_2) = (1-\rho_T) \frac{A(S_1(z_1), S_2(z_2))(z_1 - A(Y_1(S_2(z_2)), S_2(z_2)))(z_2 - 1)}{(z_1 - A(S_1(z_1), S_2(z_2)))(z_2 - A(Y_1(S_2(z_2)), S_2(z_2)))}. \tag{4.135}
$$

4.6 Packet delay

In this section, we study the steady-state packet delay. We first calculate the pgf of the packet delay of a class-1 packet, a class-2 packet and a random packet respectively. These pgf's are used to analyze the performance measures, such as the mean values and the tail probabilities.

4.6.1 Pgf $D_1(z)$ of the class-1 packet delay

We can analyze the packet delay of class-1 packets as if they are the only packets in the system. Indeed, because of the *preemptive* property of the studied priority scheduling, the class-1 packet delay is independent of class-2 packets. The pgf of the class-1 packet delay is thus the same pgf as the pgf of the delay in a single-class system with the numbers of arrivals i.i.d. from slot-to-slot with a common pgf $A_1(z)$ and general service times with pgf $S_1(z)$. This system is e.g. analyzed in Bruneel and Kim [1993] and the pgf of the packet delay of class-1 packets is given by

$$
D_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1}, \tag{4.136}
$$

with λ_1 and ρ_1 the class-1 arrival rate and class-1 arrival load respectively.

Note that this expression equals the pgf of the class-1 packet delay in a queue with an *NP* priority scheduling discipline and *deterministic class-2 service times of one slot*. Indeed, substituting $S_2(z)$ by z (and μ_2 by 1) in expression (3.115) yields expression (4.136). Although class-1 packets arriving during a class-2 service time have to wait (at least) until this class-2 service time is completed in the NP case, this service time is completed at the end of the slot in case of single-slot service times. Thus (one of) the arriving class-1 packet(s) enters the server at the beginning of the next slot. The same happens in the preemptive case (with *general* class-2 service times): class-1 packets arriving during a class-2 service time interrupt this service and one of them enters the server at the beginning of the next slot.

4.6.2 Pgf $D_2(z)$ of the class-2 packet delay

Because of the priority discipline, an expression for $D_2(z)$ will be more involved. We tag a class-2 packet that enters the buffer during slot k. We will - as in the previous chapters - use the notion of *sub-busy periods* to analyze the class-2 packet delay. Two different kinds of sub-busy periods are defined, notably, *sub-busy periods initiated by a class-1 packet* and *sub-busy periods initiated by a class-2 packet*. The first type is defined as follows: it starts at the beginning of the slot the initiating class-1 packet enters the server and ends when the number of class-1 packets in the system is one less - for the first time - as when the initiating class-1 packet entered the server. A sub-busy period initiated by a class-2 packet is defined as: it starts at the beginning of the slot the initiating class-2 packet enters the server for the first time and it ends at the beginning of which a new class-2 packet can enter the server (if there is any).

We denote the pgf of the length of a sub-busy period initiated by a class-j packet by $V_j(z)$ ($j = 1, 2$). Note that these pgf's are identical in the case of a PR and an NP priority scheduling discipline. For $V_1(z)$, this is clear: during a sub-busy period initiated by a class-1 packet only class-1 packets are served and therefore the pgf $V_1(z)$ is independent of the type of the priority scheduling discipline. A sub-busy period initiated by a class-2 packet starts with the service of a class-2 packet. The class-1 system contents is thus zero at that time instant (in both the NP and the PR case). It ends when two conditions are satisfied for the first time: firstly, the initiating class-2 packet is fully transmitted and secondly, the class-1 system contents is zero (again). Since both types of priority scheduling are work-conserving, these two conditions are satisfied at exactly the same time for both priority disciplines (if the same number of arrivals occur during the sub-busy period for both scheduling types). Therefore the lengths of the class-2 sub-busy periods are equally distributed for both priority types. The $V_j(z)$ are thus identical as in the previous chapter, namely

$$V_j(z) = S_j(zA_1(V_1(z))), \tag{4.137}$$

with $j = 1, 2$.

We now perform a similar delay analysis as in the previous chapter (subsection 3.6.2). Let us again refer to the packets in the system at the end of slot k, but that have to be served before the tagged packet as the "primary packets". So, the tagged class-2 packet enters the server for the first time when *all primary packets* and *all class-1 packets that arrived after slot k* (i.e., while the tagged packet is waiting in the queue) are served. All primary class-j packets add a class-j sub-busy period to the delay of the tagged packet. Let $\tilde{v}_{j,m}$ denote the length of the m-th class-j sub-busy period added to the tagged packet's delay by the m-th class-j packet already in the queue at the beginning of slot k and let $v_{j,m}^{(i)}$ denote the length of the sub-busy period added to the delay of the tagged class-2 packet by the m-th class-j packet that arrives during slot i.

The delay of the tagged packet depends on the state of the system at the beginning of its arrival slot (slot k by definition). This state is described by the class-1 and class-2 system contents $u_{1,k}$ and $u_{2,k}$ and the residual service times of the oldest class-1 and class-2 packet $r_{1,k}$ and $r_{2,k}$ respectively, which are defined earlier in this chapter. The delay of the tagged packet is written as a function of this state vector $\{r_{1,k}, u_{1,k}, r_{2,k}, u_{2,k}\}$ at the beginning of its arrival slot as follows:

$$\begin{aligned} d_2 =& [r_{1,k} + r_{2,k} - 1]^+ + \sum_{i=1}^{[r_{1,k}+r_{2,k}-1]^+} \sum_{m=1}^{a_{1,n_i}} v_{1,m}^{(n_i)} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} \\ &+ \sum_{j=1}^{2} \sum_{m=1}^{[u_{j,k}-1]^+} \tilde{v}_{j,m} + s_2^* + \sum_{i=1}^{s_2^*-1} \sum_{m=1}^{a_{1,l_i}} v_{1,m}^{(l_i)}, \end{aligned} \tag{4.138}$$

with $f_{j,k}$, $j = 1, 2$ defined as the numbers of class-j packets arriving during slot k, but that have to be served before the tagged packet. Slots n_i ($i = 1, \ldots, [r_{1,k} + r_{2,k} - 1]^+$) are defined as the slots that the oldest class-1 and class-2 packet receive service and the slots l_i ($i = 1, \ldots, s_2^*$) are defined as the slots during which the tagged packet receives service. Expression (4.138) of the class-2 packet delay exists of the following parts: the remaining service times of the oldest class-1 packet and class-2 packet at the beginning of slot k (note that the remaining service time of the class-2 packet can exist of non-consecutive slots, because of the preemptive property), and the sub-busy periods added by the class-1 packets arriving during these remaining service times. These contribute in the first 2 terms of (4.138). The class-j primary packets that arrive during slot k and that have to be served before the tagged packet add $f_{j,k}$ class-j sub-busy periods to the delay. These contribute in the third term. The sub-busy periods added by the class-1 and class-2 packets already in the system (excluding the oldest ones) at the beginning of slot k contribute in the fourth term. Finally the service time of the tagged class-2 packet itself and the sub-busy periods of the class-1 packets arriving during this service time (except for its last slot) contribute in the last two terms. Note that

the class-1 packets arriving during the last slot of the tagged packet's service time do not influence its delay, since the tagged packet's service is completed at the end of that slot.

We distinguish four possibilities of the system state at the beginning of slot k: $u_{1,k} = u_{2,k} = 0$; $u_{1,k} = 0, u_{2,k} > 0$; $u_{1,k} > 0, u_{2,k} = 0$ and $u_{1,k} > 0, u_{2,k} > 0$. We thus find

$$\begin{aligned} D_2(z) =& \mathrm{E}\left[z^{d_2}\{u_{1,k} = u_{2,k} = 0\}\right] + \mathrm{E}\left[z^{d_2}\{u_{1,k} = 0, u_{2,k} > 0\}\right] \\ &+ \mathrm{E}\left[z^{d_2}\{u_{1,k} > 0, u_{2,k} = 0\}\right] + \mathrm{E}\left[z^{d_2}\{u_{1,k} > 0, u_{2,k} > 0\}\right]. \end{aligned} \tag{4.139}$$

Using expression (4.138) of d_2 in this expression yields

$$\begin{aligned} D_2(z) =& \mathrm{E}\left[z^{\sum_{j=1}^{2}\sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)}}\right] \mathrm{E}\left[z^{s_2^* + \sum_{i=1}^{s_2^*-1}\sum_{m=1}^{a_{1,l_i}} v_{1,m}^{(l_i)}}\right] \Bigg\{ \mathrm{Prob}[u_{1,k} = u_{2,k} = 0] \\ &+ \mathrm{E}\left[z^{\sum_{i=1}^{r_{2,k}-1}\left(1+\sum_{m=1}^{a_{1,n_i}} v_{1,m}^{(n_i)}\right)+\sum_{m=1}^{u_{2,k}-1}\tilde{v}_{2,m}}\{u_{1,k} = 0, u_{2,k} > 0\}\right] \\ &+ \mathrm{E}\left[z^{\sum_{i=1}^{r_{1,k}-1}\left(1+\sum_{m=1}^{a_{1,n_i}} v_{1,m}^{(n_i)}\right)+\sum_{m=1}^{u_{1,k}-1}\tilde{v}_{j,m}}\{u_{1,k} > 0, u_{2,k} = 0\}\right] \\ &+ \mathrm{E}\left[z^{\left\{\begin{array}{l}\sum_{i=1}^{r_{1,k}+r_{2,k}-1}\left(1+\sum_{m=1}^{a_{1,n_i}} v_{1,m}^{(n_i)}\right) \\ +\sum_{j=1}^{2}\sum_{m=1}^{[u_{j,k}-1]^+}\tilde{v}_{j,m}\end{array}\right\}}\{u_{1,k} > 0, u_{2,k} > 0\}\right]\Bigg\}. \end{aligned} \tag{4.140}$$

The first factor equals the pgf of the total number of sub-busy periods added to the tagged packet's delay by the packets arriving in the same slot as the tagged one, but that have to be served before it. The second factor equals the pgf of the last part of the tagged packet's delay, namely, starting at the slot the tagged packet enters the server for the first time. Finally, the influence of the state of the system at the beginning of slot k is given by the third factor. This factor can thus be written in terms of the pgf $P(x_1, z_1, x_2, z_2)$. We get

$$\begin{aligned} D_2(z) =& F^{(2)}(V_1(z), V_2(z)) \frac{S_2(zA_1(V_1(z)))}{A_1(V_1(z))} \Bigg\{ P(0,0,0,0) \\ &+ \frac{P(0,0,zA_1(V_1(z)),V_2(z)) - P(0,0,0,0)}{zA_1(V_1(z))V_2(z)} \\ &+ \frac{P(zA_1(V_1(z)),V_1(z),0,0) - P(0,0,0,0)}{zA_1(V_1(z))V_1(z)} \\ &+ \frac{1}{zA_1(V_1(z))V_1(z)V_2(z)}[P(zA_1(V_1(z)),V_1(z),zA_1(V_1(z)),V_2(z)) \\ &- P(0,0,zA_1(V_1(z)),V_2(z)) - P(zA_1(V_1(z)),V_1(z),0,0) \end{aligned} \tag{4.141}$$

$$+ P(0,0,0,0)]\Bigg\},$$

with $F^{(2)}(z_1,z_2) \triangleq E[z_1^{f_{1,k}} z_2^{f_{2,k}}]$. The random variables $f_{1,k}$ and $f_{2,k}$ have the following joint pgf (for more details see chapter 3):

$$F^{(2)}(z_1,z_2) = \frac{A(z_1,z_2) - A_1(z_1)}{\lambda_2(z_2-1)}. \tag{4.142}$$

Substituting expressions (4.84) and (4.142) in expression (4.141) leads to the following final expression of $D_2(z)$:

$$D_2(z) = \frac{1-\rho_T}{\lambda_2} \frac{V_2(z)(zA_1(V_1(z))-1)}{A_1(V_1(z))(V_2(z)-1)} \frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))}, \tag{4.143}$$

with $V_j(z)$ given by (4.137), $j = 1, 2$.

4.6.3 Pgf $D(z)$ of the delay of a random packet

In this subsection, we tag a random (class-1 or class-2) packet. Since, the probability that the tagged packet is of class-j is equal to λ_j/λ_T,

$$D(z) = \frac{\lambda_1}{\lambda_T} D_1(z) + \frac{\lambda_2}{\lambda_T} D_2(z). \tag{4.144}$$

Substituting (4.136) and (4.143) in this expression yields

$$\begin{aligned} D(z) =& \frac{1-\rho_1}{\lambda_T} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z))-1}{S_1(z)-1} \\ &+ \frac{1-\rho_T}{\lambda_T} \frac{V_2(z)(zA_1(V_1(z))-1)}{A_1(V_1(z))(V_2(z)-1)} \frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))}. \end{aligned} \tag{4.145}$$

4.6.4 Calculation of moments

From the pgf's $D_1(z)$, $D_2(z)$ and $D(z)$, all moments of the delays of a class-1, of a class-2 and of a random packet can be calculated. We show the mean values in this subsection.

The mean class-1 delay is given by

$$\mathrm{E}[d_1] = D_1'(1) \tag{4.146}$$

$$= \frac{\mu_1}{2} + \frac{\mu_1 \mathrm{Var}[a_1]}{2\lambda_1(1-\rho_1)} + \frac{\lambda_1 \mathrm{Var}[s_1]}{2(1-\rho_1)}, \tag{4.147}$$

with Var$[a_1]$ and Var$[s_1]$, the variance of the number of per-slot class-1 arrivals and of the class-1 service times respectively.

Taking the first derivative of expression (4.143), the mean class-2 delay is found to be

$$\begin{aligned}
\mathrm{E}[d_2] =& D_2'(1) && (4.148)\\
=& \frac{\mu_2}{2} + \frac{\mu_1^2 \mathrm{Var}[a_1]}{2(1-\rho_T)(1-\rho_1)} + \frac{\mu_2 \mathrm{Var}[a_2]}{2\lambda_2(1-\rho_T)} + \frac{\mu_1 \mathrm{Cov}[a_1,a_2]}{\lambda_2(1-\rho_T)} && (4.149)\\
&+ \frac{\lambda_1 \mathrm{Var}[s_1] + \lambda_2 \mathrm{Var}[s_2]}{2(1-\rho_T)(1-\rho_1)} + \frac{\rho_1(\mu_2-1)}{2(1-\rho_1)}.
\end{aligned}$$

Var$[a_2]$ and Var$[s_2]$ are the variances of the number of per-slot class-2 arrivals and of the class-2 service times respectively.

Finally, the mean delay of a random packet is given by

$$\begin{aligned}
\mathrm{E}[d] =& D'(1) && (4.150)\\
=& \frac{\lambda_1}{\lambda_T}\mathrm{E}[d_1] + \frac{\lambda_2}{\lambda_T}\mathrm{E}[d_2] && (4.151)\\
=& \frac{\rho_T}{2\lambda_T} + \frac{\mu_1 \mathrm{Var}[a_T]}{2\lambda_T(1-\rho_T)} - \frac{\mu_1\lambda_2(\mu_2-\mu_1)\mathrm{Var}[a_1]}{2\lambda_T(1-\rho_T)(1-\rho_1)} + \frac{(\mu_2-\mu_1)\mathrm{Var}[a_2]}{2\lambda_T(1-\rho_T)} && (4.152)\\
&+ \frac{\lambda_1 \mathrm{Var}[s_1](\lambda_1(1-\rho_1) + \lambda_2(1-\lambda_1\mu_2))}{2\lambda_T(1-\rho_T)(1-\rho_1)} + \frac{\lambda_2^2 \mathrm{Var}[s_2]}{2\lambda_T(1-\rho_T)(1-\rho_1)}\\
&+ \frac{\rho_1\lambda_2(\mu_2-1)}{2\lambda_T(1-\rho_1)}.
\end{aligned}$$

Little's law can be seen to hold for the total system and for the classes separately (see expressions (4.106)-(4.152), (4.108)-(4.147) and (4.110)-(4.149)).

4.6.5 Calculation of tail probabilities

In this subsection, we retrieve expressions of tail probabilities of the delay of a class-1, of a class-2 and a of a random packet respectively.

Class-1 packet delay

From expression (4.136), it is seen that the dominant pole $\hat{z}_H$ of $D_1(z)$ is a zero of $z - A_1(S_1(z))$ and that this pole has multiplicity one. We thus obtain

$$\begin{aligned}
d_1(n) \triangleq& \mathrm{Prob}[d_1 = n] && (4.153)\\
\approx& \hat{K}_1 \hat{z}_H^{-n-1}, && (4.154)
\end{aligned}$$

for large enough n, and

$$\text{Prob}[d_1 > D] \approx \frac{\hat{K}_1 \hat{z}_H^{-D-1}}{\hat{z}_H - 1}, \tag{4.155}$$

with

$$\hat{K}_1 = \frac{(1-\rho_1) S_1(\hat{z}_H) \left(\hat{z}_H - 1\right)^2}{\lambda_1 \left(S_1(\hat{z}_H) - 1\right) \left(A_1'(S_1(\hat{z}_H)) S_1'(\hat{z}_H) - 1\right)}. \tag{4.156}$$

Behavior of $V_1(z)$ and $V_2(z)$

Since the $V_j(z)$ are equally defined as in chapter 3 and since we have studied their behavior in that chapter (subsection 3.6.6), we will only summarize the results in this chapter.

Both $V_j(z)$ have the same branch-point singularity $\hat{z}_B$ with

$$\hat{z}_B S_1'(\hat{z}_B A_1(V_1(\hat{z}_B))) A_1'(V_1(\hat{z}_B)) = 1. \tag{4.157}$$

In the neighborhood of $\hat{z}_B$, $V_j(z)$ behaves as

$$V_j(z) \approx V_j(\hat{z}_B) - K_{V_j} \sqrt{\hat{z}_B - z}, \tag{4.158}$$

with

$$K_{V_1} = \sqrt{\frac{2A_1(V_1(\hat{z}_B))}{\hat{z}_B A_1''(V_1(\hat{z}_B)) + S_1''(\hat{z}_B A_1(V_1(\hat{z}_B)))(\hat{z}_B A_1'(V_1(\hat{z}_B)))^3}}, \tag{4.159}$$

and

$$K_{V_2} = K_{V_1} S_2'(\hat{z}_B A_1(V_1(\hat{z}_B))) \hat{z}_B A_1'(V_1(\hat{z}_B)). \tag{4.160}$$

Class-2 packet delay

Expression (4.143) of $D_2(z)$ has two important singularities: the branch point $\hat{z}_B$ and a pole $\hat{z}_L$ with multiplicity 1 which is a zero of $zA_1(V_1(z)) - A(V_1(z), V_2(z))$. Both singularities can be dominant, depending on the input parameters/pgf's of the queueing system. The tail behavior of the class-2 packet delay is summarized as:

$$d_2(n) \triangleq \text{Prob}[d_2 = n] \tag{4.161}$$

$$\approx \begin{cases} \hat{K}_2^{(1)} \hat{z}_L^{-n-1} & \text{if } \hat{z}_L \text{ dominant} \\ \dfrac{\hat{K}_2^{(2)} n^{-1/2} \hat{z}_B^{-n}}{\sqrt{\hat{z}_B \pi}} & \text{if } \hat{z}_L = \hat{z}_B \text{ dominant} \\ \dfrac{\hat{K}_2^{(3)}}{2} \sqrt{\dfrac{\hat{z}_B}{\pi}} n^{-3/2} \hat{z}_B^{-n} & \text{if } \hat{z}_B \text{ dominant,} \end{cases} \tag{4.162}$$

and

$$\text{Prob}[d_2 > D] \approx \frac{d_2(D)}{\hat{z}_* - 1}, \tag{4.163}$$

with $\hat{z}_*$ the dominant singularity of $D_2(z)$ and with

$$\hat{K}_2^{(1)} = \frac{(1-\rho_T) V_2(\hat{z}_L)(\hat{z}_L - 1)(\hat{z}_L A_1(V_1(\hat{z}_L)) - 1)}{\lambda_2 (V_2(\hat{z}_L) - 1) \left(\dfrac{dA(V_1(z), V_2(z))}{dz} - \dfrac{d(zA_1(V_1(z))}{dz} \right) \Bigg|_{z=\hat{z}_L}} \tag{4.164}$$

$$\hat{K}_2^{(2)} = \frac{(1-\rho_T)(\hat{z}_B A_1(V_1(\hat{z}_B)) - 1)}{\lambda_2 (V_2(\hat{z}_B) - 1)} \tag{4.165}$$

$$\times \frac{V_2(\hat{z}_B)(\hat{z}_B - 1)}{K_{V_1} \left(A^{(1)}(V_1(\hat{z}_B), V_2(\hat{z}_B)) - \hat{z}_B A_1'(V_1(\hat{z}_B)) \right) + K_{V_2} A^{(2)}(V_1(\hat{z}_B), V_2(\hat{z}_B))}$$

$$\hat{K}_2^{(3)} = \frac{1-\rho_T}{\lambda_2 A_1(V_1(\hat{z}_B))^2 (V_2(\hat{z}_B) - 1)^2 (\hat{z}_B A_1(V_1(\hat{z}_B)) - A(V_1(\hat{z}_B), V_2(\hat{z}_B)))^2} \tag{4.166}$$

$$\times \Big\{ \Big[A_1(V_1(\hat{z}_B)) \left(K_{V_1} A^{(1)}(V_1(\hat{z}_B), V_2(\hat{z}_B)) + K_{V_2} A^{(2)}(V_1(\hat{z}_B), V_2(\hat{z}_B)) \right)$$

$$- K_{V_1} A(V_1(\hat{z}_B), V_2(\hat{z}_B)) A_1'(V_1(\hat{z}_B)) \Big] (\hat{z}_B - 1)(\hat{z}_B A_1(V_1(\hat{z}_B)) - 1)$$

$$\times (V_2(\hat{z}_B) - 1) A_1(V_1(\hat{z}_B)) V_2(\hat{z}_B) + [K_{V_1} V_2(\hat{z}_B) A_1'(V_1(\hat{z}_B))(V_2(\hat{z}_B) - 1)$$

$$- K_{V_2} A_1(V_1(\hat{z}_B))(\hat{z}_B A_1(V_1(\hat{z}_B)) - 1)] \left(A(V_1(\hat{z}_B), V_2(\hat{z}_B)) - A_1(V_1(\hat{z}_B)) \right)$$

$$\times \left(\hat{z}_B A_1(V_1(\hat{z}_B)) - A(V_1(\hat{z}_B), V_2(\hat{z}_B)) \right) \Big\}. \tag{4.167}$$

Delay random packet

Finally, we calculate the tail probabilities of the delay of a random packet. Its pgf $D(z)$ is given by expression (4.145). The dominant singularity of this function is $\hat{z}_L$ or $\hat{z}_B$, depending on which is smallest.

Note that - similar as in the previous chapter - the dominant singularity of $D_1(z)$ - $\hat{z}_H$ - is also a singularity of $D(z)$, but that $\hat{z}_H$ is never dominant, since $\hat{z}_H > \hat{z}_B$. This inequality is proved in section 3.6.6. From expression (4.144) and the fact that the dominant singularities of $D(z)$ and $D_2(z)$ are equal, it is

easily seen that

$$d(n) \triangleq \text{Prob}[d = n] \tag{4.168}$$

$$\approx \frac{\lambda_2}{\lambda_T} d_2(n), \tag{4.169}$$

for large enough n. Since $d_2(n)$ is approximately calculated in expression (4.162), $d(n)$ is approximately determined from this expression.

Finally, the probability that the steady-state delay of a random packet is larger than a bound D is given by

$$\text{Prob}[d > D] \approx \frac{\lambda_2}{\lambda_T} \text{Prob}[d_2 > D]. \tag{4.170}$$

4.7 Waiting time

The waiting time is defined as the number of slots a packet has to wait in the *queue* before starting service. Thus specifically for the class-2 packets, the waiting time - as defined in this dissertation - does not include the slots the packets spend in the queue after the possible interruption.

The relation between the waiting time t_1 and the delay d_1 of a class-1 packet is given by

$$d_1 = t_1 + s_1^*, \tag{4.171}$$

with s_1^* the service time of the packet. Since t_1 and s_1^* are independent variables, we find

$$T_1(z) = \frac{D_1(z)}{S_1(z)}, \tag{4.172}$$

for the pgf of the steady-state class-1 waiting time. Substituting expression (4.136) in this expression yields

$$T_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{z-1}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1}, \tag{4.173}$$

for the pgf of the class-1 waiting time.

The relation between the waiting time t_2 and the delay d_2 of a class-2 packet is given by

$$d_2 = t_2 + s_2^* + \sum_{i=1}^{s_2^*-1} \sum_{m=1}^{a_{1,l_i}} v_{1,m}^{(l_i)}, \tag{4.174}$$

with s_2^* the service time of the packet and the $v_{1,m}^{(l_i)}$ the sub-busy periods added to the delay by the class-1 packets arriving during the service slots of the class-2 packet (see also expression (4.138)). Since the first term of the right-hand side of expression (4.174) is independent of the other two terms, we find

$$T_2(z) = D_2(z)\frac{A_1(V_1(z))}{S_2(zA_1(V_1(z)))}, \tag{4.175}$$

for the pgf of the class-2 waiting time. Substituting expression (4.143) in this expression gives

$$T_2(z) = \frac{1-\rho_T}{\lambda_2}\frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))}\frac{zA_1(V_1(z))-1}{V_2(z)-1}. \tag{4.176}$$

Finally, the pgf of the steady-state waiting time of a random (class-1 or class-2) packet is given by

$$T(z) = \frac{\lambda_1}{\lambda_T}T_1(z) + \frac{\lambda_2}{\lambda_T}T_2(z) \tag{4.177}$$

$$\begin{aligned} &= \frac{1-\rho_1}{\lambda_T}\frac{z-1}{z-A_1(S_1(z))}\frac{A_1(S_1(z))-1}{S_1(z)-1} \\ &\quad + \frac{1-\rho_T}{\lambda_T}\frac{A(V_1(z),V_2(z)) - A_1(V_1(z))}{zA_1(V_1(z)) - A(V_1(z),V_2(z))}\frac{zA_1(V_1(z))-1}{V_2(z)-1}. \end{aligned} \tag{4.178}$$

4.8 Identical variables in PR and NP priority queues

There exist some stochastic variables in case of an NP (studied in section 3) and a PR (studied in this section) priority queue which are identically distributed.

4.8.1 Total unfinished work

The total unfinished work at the beginning of a random slot is independent of the scheduling discipline, as long as this scheduling discipline is work-conserving. Since both the NP and PR priority scheduling disciplines are work-conserving disciplines, the total unfinished work is equally distributed for both scheduling disciplines. Indeed, using expressions (3.103) and (4.135) of the joint pgf of the unfinished work of both classes in case of the NP and PR priority scheduling discipline respectively, we find

$$W_T(z) = W(z,z) \tag{4.179}$$

$$= (1-\rho_T)\frac{A(S_1(z),S_2(z))(z-1)}{z-A(S_1(z),S_2(z))}, \tag{4.180}$$

for the pgf of the total unfinished work in both cases. More generally, this expression equals the pgf of the (total) unfinished work in a queueing system where the pgf of the work arriving in the system during a slot is equal to $A(S_1(z), S_2(z))$.

4.8.2 Class-2 waiting time

The class-2 waiting times in queues with the NP or the PR priority discipline are closely related to the total unfinished work in these queues. In this subsection, we show an alternative calculation of the pgf $T_2(z)$ of the class-2 waiting time, based on expression (4.180) of $W_T(z)$. Furthermore, we show that the class-2 waiting times in a PR and an NP priority queue are equally distributed.

Denoting t_2 as the waiting time of a (tagged) class-2 packet, w_T the total unfinished work at the beginning of this packet's arrival slot and g_T the total amount of *work* arriving in the same slot as and ultimately served before the tagged packet, t_2 is given by

$$t_2 = \sum_{i=1}^{[w_T-1]^+ + g_T} v_i, \tag{4.181}$$

where the v_i are defined as the number of slots necessary to lower the work ahead of the tagged class-2 packet by 1 (taking into account newly arriving class-1 packets). All v_i are i.i.d. and their common pgf is given by

$$V(z) = zA_1(V_1(z)), \tag{4.182}$$

since the v_i consist of one slot augmented with the sub-busy periods added by the class-1 arrivals during that slot. Recall that

$$V_1(z) = S_1(zA_1(V_1(z))). \tag{4.183}$$

Since w_T and g_T are independent stochastic variables, z-transforming (4.181) yields

$$T_2(z) = \frac{(V(z)-1)W_T(0) + W_T(V(z))}{V(z)} G_T(V(z)), \tag{4.184}$$

with $W_T(z)$ and $G_T(z)$ the pgf of w_T and g_T respectively. $W_T(z)$ is given by expression (4.180) and $G_T(z)$ is given by

$$G_T(z) = F^{(2)}(S_1(z), S_2(z)), \tag{4.185}$$

with $F^{(2)}(z_1, z_2)$ given by expression (4.142). Substituting all this in expression (4.184) finally yields expressions (3.166) and (4.176).

Concluding, (the pgf of) the class-2 waiting time is identical for both the NP and PR priority scheduling discipline, because in both cases a class-2 packet can only start service when all class-2 packets arrived before it and *all* class-1 packets are serviced. Whether or not the class-1 packets interrupt class-2 service times is not important for the class-2 waiting time.

Finally, we note that the reasoning in this section is part of the approach followed in [Takahashi and Hashida 1991] to analyze the packet delays in NP and PR priority queues.

4.9 Numerical examples

In this section, we discuss some numerical examples. We have studied the influence of different parameters on system contents and packet delay performance measures in the case of an NP priority scheduling discipline quite extensively (in the previous chapter). Furthermore, the general influence of these parameters is often similar for the NP or the PR priority scheduling disciplines. Therefore, we focus - in this section - mostly on the comparison between the results for these priority disciplines.

4.9.1 Input processes

We first briefly summarize the most important characteristics of the arrival and service processes we consider in this section.

The arrival process

The pgf of the number of per-slot class-1 and class-2 arrivals is given by

$$A(z_1, z_2) = \left(1 - \frac{\lambda_1}{N}(1 - z_1) - \frac{\lambda_2}{N}(1 - z_2)\right)^N. \tag{4.186}$$

N is chosen 16 in the figures in this section.

The means of the total, class-1 and class-2 number of per-slot arrivals are given by λ_T, λ_1 and λ_2 respectively.

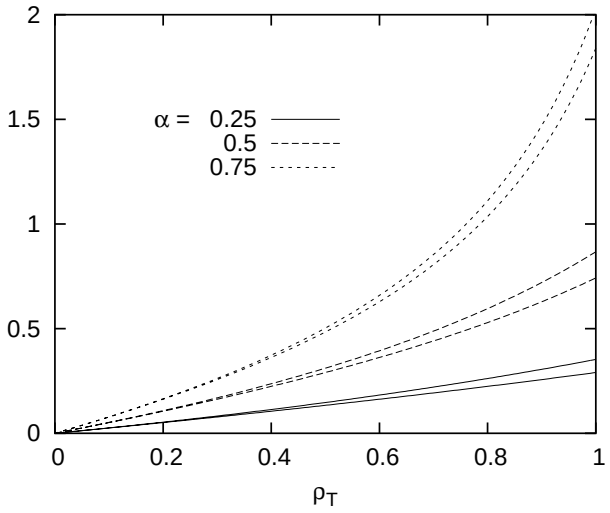


Figure 4.3: Mean class-1 system contents versus the total arrival rate for both the PR (lower curves) as the NP priority (upper curves) scheduling disciplines ($\mu_1 = \mu_2 = 2$)

The service process

The service times of both classes are assumed deterministic throughout this section, i.e.,

$$S_j(z) = z^{\mu_j}, \tag{4.187}$$

$j = 1, 2$, with μ_j the class-j service time.

4.9.2 Influence of the load

System contents

In Figures 4.3 and 4.4, we show the mean class-1 and class-2 system contents respectively as functions of the total load, with $\mu_1 = \mu_2 = 2$ and with $\alpha \triangleq \rho_1/\rho_T$ - equal to 0.25, 0.5 and 0.75. In both figures, we show the curves for both the PR and the NP priority scheduling disciplines. The mean class-1 system contents is larger in case of the NP priority scheduling. The opposite holds for the mean class-2 system contents. It is seen that the difference between both scheduling disciplines is (relatively) more significant for the mean class-1 system contents, especially as $\rho_T \to 1$.

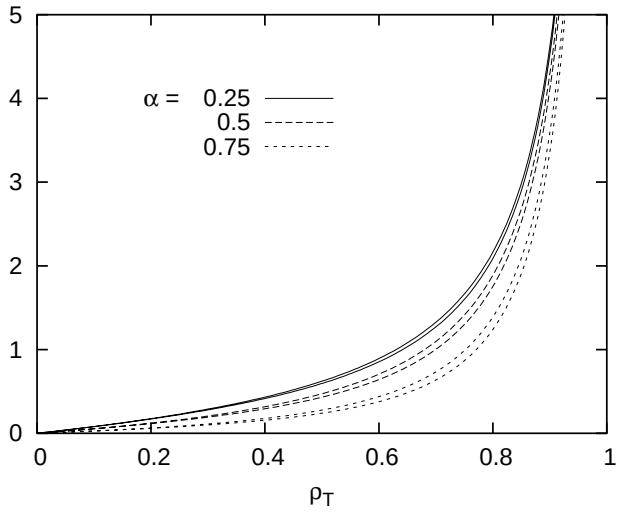


Figure 4.4: Mean class-2 system contents versus the total arrival rate for both the PR (upper curves) as the NP priority (lower curves) scheduling disciplines ($\mu_1 = \mu_2 = 2$)

Packet delay

Similar differences between the PR and the NP priority scheduling disciplines can be seen for the mean packet delay. Figures 4.5 and 4.6 depict the mean class-1 and class-2 packet delays respectively as functions of the total load, with $\mu_1 = \mu_2 = 2$ and with α equal to 0.25, 0.5 and 0.75. It can be seen that the mean class-1 delay can be considerably higher in the NP priority case, especially for low α. On the other hand for high α, the mean delay of the class-2 packets is significantly higher in the case of the PR priority scheduling. Indeed, for high α, a service time of a class-2 packet is interrupted with a large probability when a PR priority scheduling is applied.

4.9.3 Influence of the service times

System contents

Figures 4.7 and 4.8 show the mean class-1 and class-2 system contents respectively versus the (deterministic) lengths of the class-1 service times for both the NP and PR priority scheduling disciplines. The total load is 0.75, the class-2 service times are deterministically equal to 20 and $\alpha = 0.25$, 0.5 and 0.75 respectively. From Figure 4.7 it is seen that the mean class-1 system contents stay constant in the case of the PR priority scheduling (for this particular arrival and service process). For low μ_1 the difference between the mean class-1

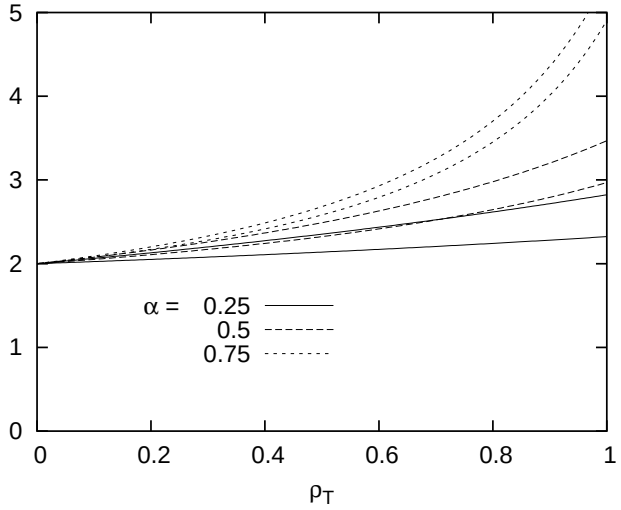


Figure 4.5: Mean class-1 packet delay versus the total arrival rate for both the PR (lower curves) as the NP priority (upper curves) scheduling disciplines ($\mu_1 = \mu_2 = 2$)

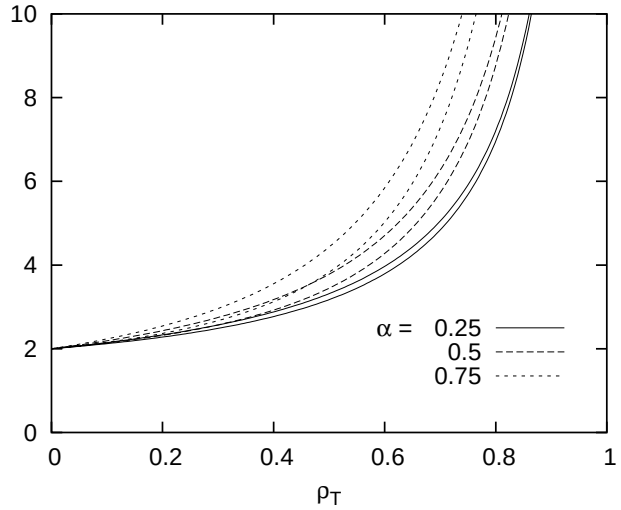


Figure 4.6: Mean class-2 packet delay versus the total arrival rate for both the PR (upper curves) as the NP priority (lower curves) scheduling disciplines ($\mu_1 = \mu_2 = 2$)

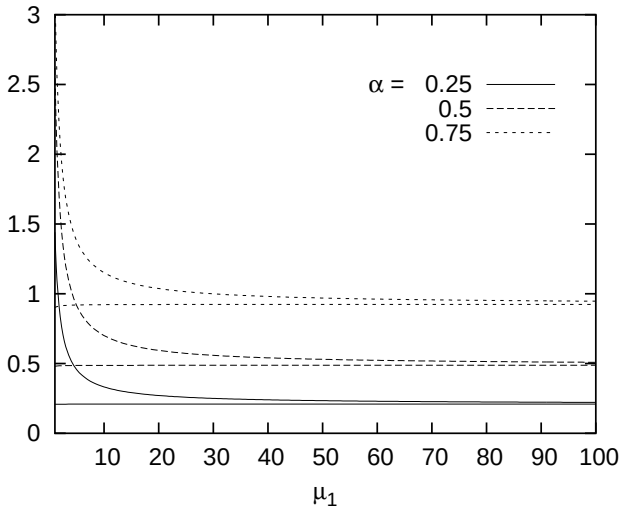


Figure 4.7: Mean class-1 system contents versus the mean class-1 service time for both the PR (lower curves) as the NP priority (upper curves) scheduling disciplines ($\rho_T = 0.75, \mu_2 = 20$)

system contents in the NP and PR case is rather large, while the difference is negligible for high μ_1. Indeed, for low μ_1 the mean class-1 system contents in the NP priority queue is largely due to a residual class-2 service time. For high μ_1, a residual service time of class-2 packets is negligible to the (residual) service times of class-1 packets, and thus the curves for the NP and PR priority queue lie near to each other for high μ_1. From Figure 4.8, we conclude that the difference between the mean class-2 system contents in the case of the PR and the NP priority queue is independent of μ_1 (when ρ_1 is kept constant).

In Figures 4.9 and 4.10, we show the mean class-1 and class-2 system contents respectively as functions of the length of the class-2 service times, with $\rho_T = 0.75$, $\mu_1 = 20$ and $\alpha = 0.25$, 0.5 and 0.75. It is seen that the mean class-1 contents are independent of the class-2 service times in case of the PR priority scheduling, as expected. For $\mu_2 = 1$, the mean class-1 system contents in the NP priority queue is equal to the mean class-1 system contents in the PR priority queue - for single-slot class-2 service times, the class-1 system contents are indeed independent of the class-2 characteristics - and the mean class-2 system contents increase with increasing μ_2.

Packet delay

Figures 4.11 and 4.12 depict the mean class-1 and class-2 packet delays respectively as functions of the mean class-1 service times for both the NP and

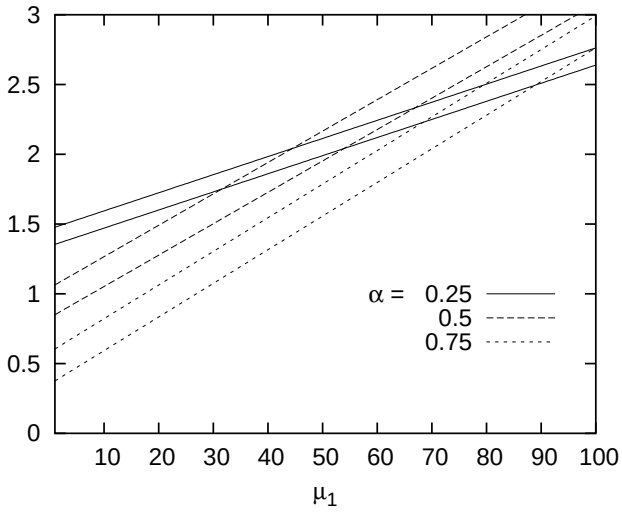


Figure 4.8: Mean class-2 system contents versus the mean class-1 service time for both the PR (upper curves) as the NP priority (lower curves) scheduling disciplines ($\rho_T = 0.75, \mu_2 = 20$)

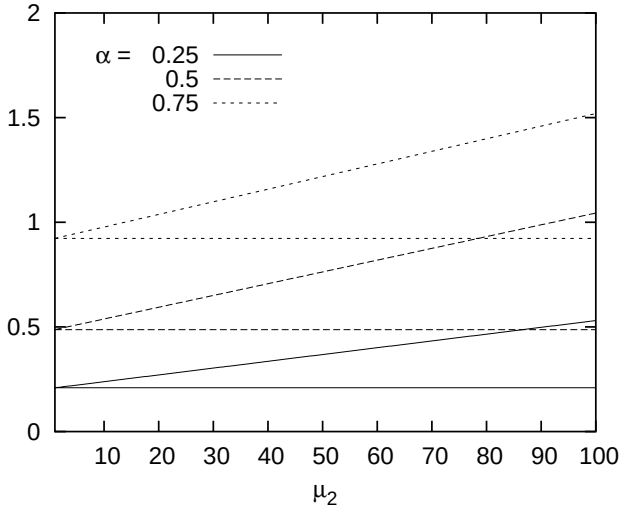


Figure 4.9: Mean class-1 system contents versus the mean class-2 service time for both the PR (lower curves) as the NP priority (upper curves) scheduling disciplines ($\rho_T = 0.75, \mu_1 = 20$)

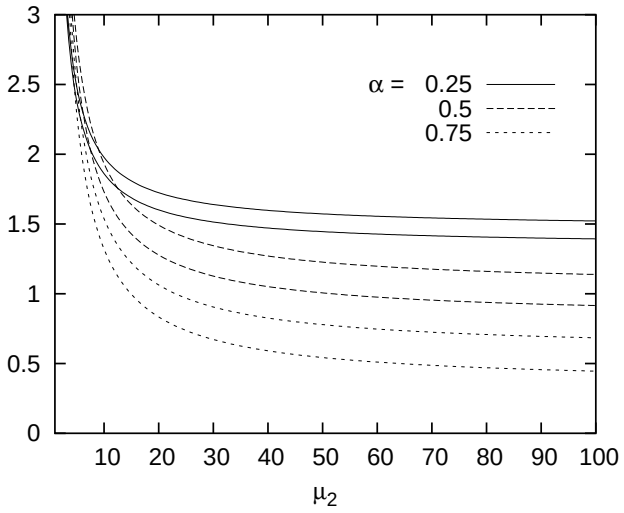


Figure 4.10: Mean class-2 system contents versus the mean class-2 service time for both the PR (upper curves) as the NP priority (lower curves) scheduling disciplines ($\rho_T = 0.75, \mu_1 = 20$)

PR priority scheduling disciplines. The total load is 0.75, the class-2 service times are equal to 20 and $\alpha = 0.25$, 0.5 and 0.75 respectively. It is seen that the difference between the mean class-1 packet delay for both scheduling disciplines is larger for smaller α. For the class-2 packet delay, the difference is larger for increasing α. It is also seen from both plots that the difference between the mean packet delays in both priority queues is independent of μ_1. So for high μ_1 (compared with μ_2), the difference between the mean delays of both types in the PR and NP priority queue is (relatively) negligible, while this is not the case for small μ_1.

In Figures 4.13 and 4.14, we show the mean class-1 and class-2 packet delays respectively as functions of the mean class-2 service time, with $\rho_T = 0.75$, $\mu_1 = 20$ and $\alpha = 0.25$, 0.5 and 0.75. Since the class-1 performance measures are independent of the class-2 characteristics in the PR priority queue, the mean class-1 packet delay is independent of the class-2 service times in case of the PR priority scheduling. For $\mu_2 = 1$, the mean class-1 packet delay in the NP priority queue is equal to the mean class-1 delay in the PR priority queue and the mean class-2 packet delay increases with increasing μ_2. The mean class-2 delay is equal for both types of scheduling disciplines for $\mu_2 = 1$. For $\mu_2 > 1$, the class-2 packet delay is larger in the case of the PR priority scheduling and the difference increases for increasing μ_2 and/or increasing α. As can be seen from Figure 4.14, the difference can be substantial for long class-2 service times and/or a high fraction of class-1 traffic.

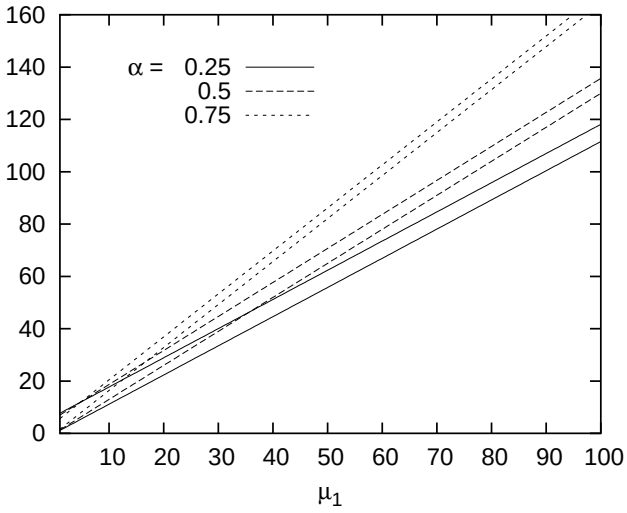


Figure 4.11: Mean class-1 packet delay versus the mean class-1 service time for both the PR (lower curves) as the NP priority (upper curves) scheduling disciplines ($\rho_T = 0.75, \mu_2 = 20$)

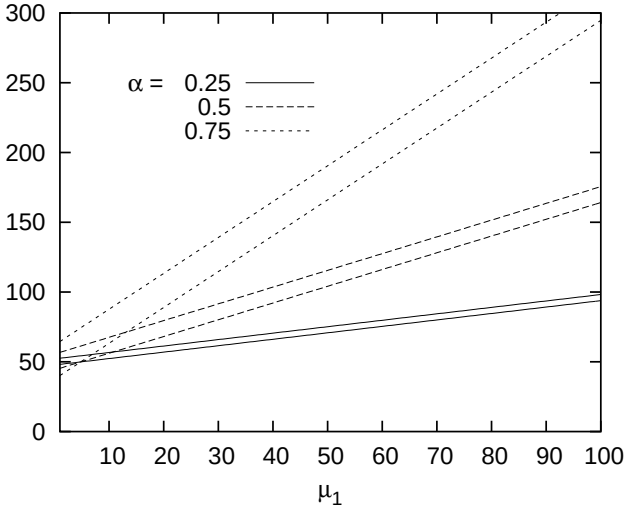


Figure 4.12: Mean class-2 packet delay versus the mean class-1 service time for both the PR (upper curves) as the NP priority (lower curves) scheduling disciplines ($\rho_T = 0.75, \mu_2 = 20$)

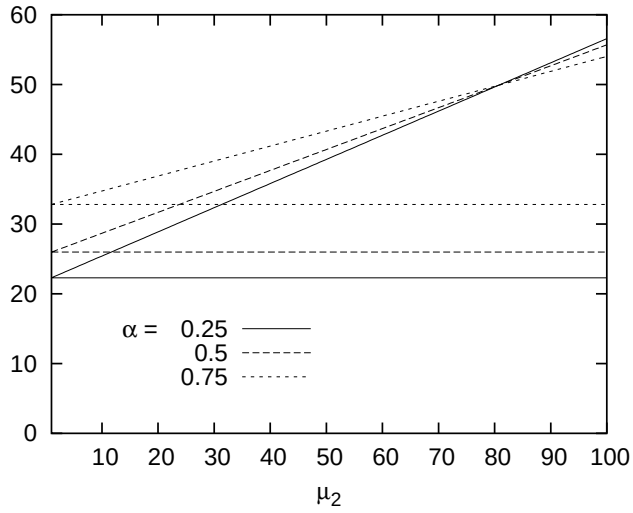


Figure 4.13: Mean class-1 packet delay versus the mean class-2 service time for both the PR (lower curves) as the NP priority (upper curves) scheduling disciplines ($\rho_T = 0.75, \mu_1 = 20$)

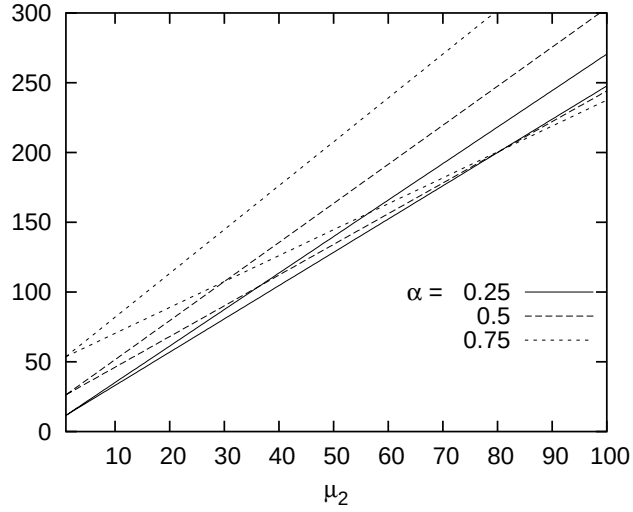


Figure 4.14: Mean class-2 packet delay versus the mean class-2 service time for both the PR (upper curves) as the NP priority (lower curves) scheduling disciplines ($\rho_T = 0.75, \mu_1 = 20$)

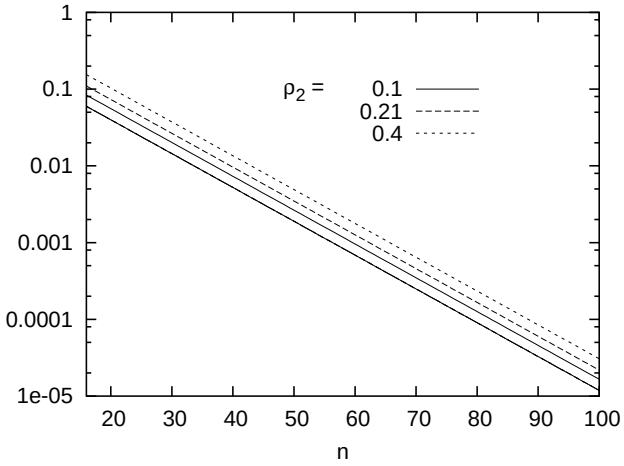


Figure 4.15: Tail probabilities of the class-1 packet delay for several class-2 loads for both the PR (lower curve) and the NP (upper curves) priority disciplines ($\rho_1 = 0.4, \mu_1 = \mu_2 = 16$)

4.9.4 Tail probabilities

In this subsection, we compare the tail behaviors of the class-1 and class-2 packet delays in case of the NP and the PR priority scheduling disciplines. We have seen that all singularities of $D_1(z)$ and $D_2(z)$ that play a role in the tail behavior are equal in case of both scheduling types. This means that the slopes of the tail probabilities will be equal for both types. In Figures 4.15 and 4.16, the tail probabilities of the class-1 and class-2 packet delay respectively are shown in case of the NP and the PR priority scheduling disciplines. The class-1 load equals 0.4 for all curves and the class-2 loads take the values 0.1 (non-geometric class-2 tail), approximately 0.21 (transition type class-2 tail) and 0.4 (geometric class-2 tail). The service times are equal to 16. In Figure 4.15, the lower curve is the curve of the tail probabilities of the class-1 packet delay in the PR priority queue. This packet delay is independent of the class-2 characteristics, and we thus get the same curve for the three chosen values of ρ_2. It is seen from these figures that the type of the priority scheduling discipline plays a significant role in the tail probabilities of both the class-1 and class-2 packet delay.

4.10 Concluding remarks

In this chapter, we analyzed a PR priority queue. In the most general case - general service times for both classes - the most important step was to find -

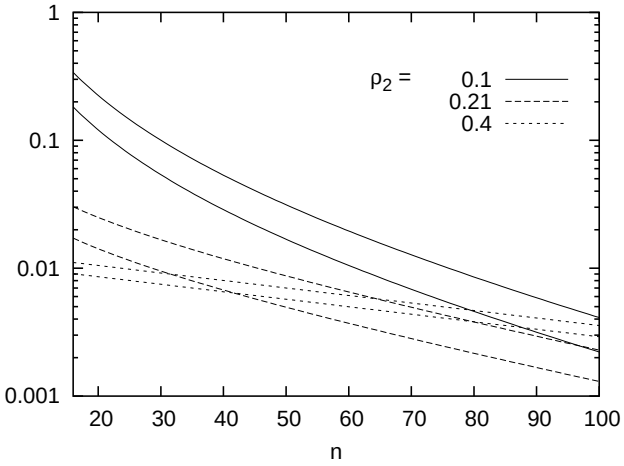


Figure 4.16: Tail probabilities of the class-2 packet delay for several class-2 loads for both the PR (upper curves) and the NP (lower curves) priority disciplines ($\rho_1 = 0.4, \mu_1 = \mu_2 = 16$)

besides the system contents of both classes - supplementary variables in order to construct a Markov chain. A four-dimensional pgf was calculated, which was the starting-point to obtain all performance measures.

In the previous chapter, a Markov chain was introduced by studying the system first at specific slot boundaries. The methods used to analyze (discrete-time) queues with general service times - namely, the method discussed in the previous chapter and the supplementary variable technique studied in this chapter - both have their advantages and disadvantages. The advantage of the analysis on specific slot boundaries is that the initial analysis is more simple, i.e., finding the joint pgf of the steady-state system contents at the specific slot boundaries is a quite straight-forward extension of the single-slot analysis. The disadvantage of this method is the quite extensive calculations which are necessary to calculate the pgf's of the required stochastic variables (system contents at random slot boundaries, unfinished work, packet delay, ...). A second disadvantage is the fact that it is sometimes hard to define specific slot boundaries so that the system contents at these slot boundaries form a Markovian description of the system. This is especially hard when service times can be interrupted (e.g., the scheduling discipline in this section). The advantages of the supplementary variable technique is the ease of defining supplementary variables to form a Markov chain and the relative ease to find the interesting pgf's once the pgf of the stochastic variables of the Markov chain is calculated. The main disadvantage is the complexity in calculating the multivariate pgf.

Comparing the performance measures - we concentrate on the (mean) packet

delay - in the *NP* and *PR* priority queue, it is seen that the NP scheduling discipline is more tolerant to the class-2 packet delay than the PR priority scheduling discipline. However, the latter discipline gives rise to a lower class-1 packet delay. The class-1 delay characteristics are completely independent of the characteristics of the class-2 traffic in the PR priority queue. This latter feature is not the case in the NP priority queue. Thus, purely reasoning on the performance measures, the PR priority scheduling discipline seems to be a better candidate to be used than the NP priority scheduling discipline since the primordial goal of a priority scheduling is to decrease the high-priority - or class-1 - packet delay. However - in e.g. telecommunication systems - the implementation of the PR priority discipline is more cumbersome, since service of packets can be interrupted and have to be (partially) restored in the queue. Even more importantly, the network will have to be able to cope with partial packets floating through the network instead of complete packets. This gives a more complicated header creation (each part of the packet its own header?), packets that have to be re-assembled, ... Since the NP priority scheduling discipline does not have these problems, this scheduling is mostly used in telecommunication systems. However, we have shown that the influence of the class-2 characteristics on the class-1 delay can be significant for this scheduling discipline, especially for long class-2 service times.

Chapter 5

Preemptive repeat priority

In this chapter, we describe the analysis of a *preemptive repeat* priority queue. Packets of two classes (class-1 and class-2) arrive in a single-server queueing system and the packets of class-1 are scheduled for service with priority over class-2 packets. So, when the server becomes available, a class-1 packet is served next (if any). If no class-1 packets are present, a class-2 packet starts service (if any). The scheduling type is *preemptive* - as in the previous chapter - which means that newly arriving class-1 packets interrupt an on-going service of a class-2 packet. Or in other words, when class-1 packets arrive during a service slot of a class-2 packet, one of the class-1 packet starts service the next slot, and the interrupted class-2 packet is pushed back in the queue (unless it was its last slot of service). In this chapter, we analyze the preemptive *repeat* priority scheduling discipline, which means that an interrupted class-2 packet has to repeat its complete service upon returning in the server (after all class-1 packets have left the system).

In the literature, two main types of preemptive repeat priority disciplines are distinguished, namely, preemptive repeat *different* (PRD) - or preemptive repeat *with resampling* - and preemptive repeat *identical* (PRI) - or *without resampling*. In the first type, an interrupted service time is resampled. Thus, when an interrupted packet enters the server for a new service attempt, the new service time is not necessarily the same as the old one, but it takes a new sample (with the same distribution). In the PRI queue, the service time of a particular class-2 packet is the same in all its service attempts. In this chapter, we study both these priority types.

Although this type of priority scheduling is mentioned - not analyzed - in earlier works like [White and Christie 1958, Miller 1960] and in e.g. the standard work of Kleinrock [1976], analyses of queues with this type of scheduling discipline are more scarce than analyses considering NP and PR priority queues. In our opinion, this is due to two (equally important) reasons. First, this type of priority scheduling is less useful in practice. This is basically be-

cause it is a *non-work-conserving* scheduling discipline, since an interrupted packet has to be fully retransmitted. In other words, the load incorporates not only the service time of class-2 packets, but also their possible repeats. The *load offered to the system* is thus different from the *arrival load*. Therefore the performance of this kind of queue is generally worse than the ones analyzed in the two previous chapters. A second reason that this kind of queue is less studied, is that its analysis is more complicated than that of the NP and PR priority queues. We go in further detail about the nature of this complexity in the first section of this chapter.

Continuous-time queues with a preemptive repeat priority discipline are analyzed by a.o. Sumita and Sheng [1988], Yoon and Un [1994] and Krinik et al. [2002]. Sumita and Sheng [1988] analyze a PRD queue in the context of a database system. The arrival process is assumed to be a Poisson process, while the service times are generally distributed. In this database system it is assumed that update requests and read queries are performed, where the former have preemptive repeat priority over the latter. The reason why the low-priority service times are resampled is because a file copy is likely to be stored on disk, and processing of update requests may change the location of the read/write heads, thus leading to a possibly resampled version of the service time of the (interrupted) read query. The authors state that therefore resampling gives a more accurate model than the model without resampling. The use of a PRI priority scheduling discipline in CSMA-CD protocols for fiber optic bus networks is described in [Yoon and Un 1994]. A general finite number (M) of stations, each with an infinite queueing capacity, are connected by an optic bus network. A station has priority of accessing the bus network over its downstream stations possibly overwriting information of the downstream stations. Therefore, this is modeled by a PRI priority queue with M priority classes. The arrival processes are modeled by a Poisson process. Finally, Krinik et al. [2002] analyze the transient behavior of a PRI priority queue with a Poisson arrival process and exponential service times using the randomization solution form and lattice path combinatorics.

In [Mukherjee et al. 1995, Fiems et al. 2004, Fiems 2004], *discrete-time* queues with a preemptive repeat priority scheduling discipline and *without correlation between the arrival processes of different classes* are studied. Mukherjee et al. [1995] study a preemptive repeat protocol for voice-data integration in a ring-based LAN. A number of stations is connected by a ring network and each station is either a voice or a data station. Voice stations can overwrite the information of the data stations. The data stations can only put data on the network when no information of the other stations is passing by. When the packet reaches its destination an acknowledgement is immediately send to the sender. The ring is unidirectional and the sender thus waits for a deterministic amount of slots - the round-trip-time (RTT) - for the acknowledgement. When no acknowledgement is received, the packet has to be retransmitted. This system is thus modeled as a preemptive repeat priority queueing system with deterministic service times (a number of slots so that the lengths of the

service times equal the RTT) for the low-priority packets. Note that the data and voice queues are all analyzed separately and that the queueing model for the low-priority queue is given by a queueing model with *service interruptions*. Fiems et al. [2004], Fiems [2004] analyze preemptive priority queues. This is also done by means of queues with service interruptions. A technique - the technique of the *effective service times* - is proposed for analyzing PR, PRD and PRI queues commonly. The interruptions are incorporated in the service times of the packets. These effective service times are obtained for the three priority disciplines separately, but once these are calculated a common queueing analysis is performed.

Walraevens et al. [2003a] analyze a *discrete-time* two-class PRD queue *with correlation between the arrival processes of the two classes*. The numbers of arrivals are i.i.d. from slot-to-slot and the service times are generally distributed. Using pgf's, the moments of the system contents and packet delay of both classes are obtained.

In this chapter, we analyze the PRD and PRI priority queues. Firstly, we start by focusing on the particular difficulties in their analyses in section 5.1. The joint pgf of the numbers of per-slot class-1 and class-2 arrivals is still given by $A(z_1, z_2)$. The service times of class-j packets are assumed to be generally distributed with their pgf's equal to $S_j(z)$, $j = 1, 2$. In this chapter, we describe the analysis of both the preemptive repeat priority queues jointly when possible and separately when necessary. The analysis of the PRD priority queue is as described in [Walraevens et al. 2003a], while the analysis of the PRI priority queue is not previously published. We make use of the supplementary variable technique - as in the previous chapter - and describe the basic analysis in chapter 5.2. The analysis of the system contents is described in section 5.3. The joint pgf's of queue contents and unfinished work of both classes are calculated in sections 5.4 and 5.5 respectively. The analyses of the packet delays and waiting times of both classes are described in sections 5.6 and 5.7 respectively. We show some numerical examples in section 5.8, where we specifically focus on the comparison of the class-2 performance measures of the PR, of the PRD and of the PRI priority queues. Finally, some concluding remarks are given in section 5.9.

5.1 Preliminaries

In this chapter, we use the *supplementary variable technique* to analyze the preemptive repeat priority queues. For more details about this technique, we refer to section 4.1.

As mentioned in the introductory paragraphs of this chapter, one of the reasons that these types of priority queues - and especially the PRI priority queues - have not been analyzed frequently in the literature is the fact that the analyses are (much) more complicated than the analyses of their NP and PR coun-

terparts. This is mainly because one has to keep track of the complete service time of the oldest class-2 packet in the queue in the PRI priority queue. (The oldest class-j packet in the system at a certain time instant is defined as the class-j packet that - of all the class-j packets present in the system at that time instant - arrived first.) This is necessary, since this packet's service can be interrupted and afterwards its complete service has to be repeated. This complexity does not occur for the PRD priority scheduling discipline, since the service time is resampled after an interruption. Therefore, the latter scheduling discipline can (also) be seen as a simplified version of the former - in terms of the analysis. We will however show that the performance measures of both queues can be considerably different.

Note that because of the more complex analysis the (finally) obtained pgf's will also be more complex. For example for the PRI priority queue an infinite sum will still appear in the pgf's of the class-2 output stochastic variables. Although means and variances are still (relatively) easy to obtain from these pgf's, calculating higher moments is in practice much more difficult. Furthermore, which singularities of the pgf's can be dominant and what the behavior of these pgf's is in the neighborhood of these singularities, is an open issue at this time. It will thus take further study to calculate tail probabilities from these pgf's.

5.2 The supplementary variable technique

We denote the system contents of class-j packets at the beginning of slot k by $u_{j,k}$ ($j = 1, 2$), as usual. The set $\{(u_{1,k}, u_{2,k}), k \geq 1\}$ does not form a Markov-chain in the case of general service times for both classes. Therefore, we introduce some additional stochastic variables.

Firstly, we define r_k as follows: r_k indicates the residual service time at the beginning of slot k, i.e., the remaining number of slots needed to serve the packet in service from the beginning of slot k on, if $u_{T,k} > 0$, and $r_k = 0$ if $u_{T,k} = 0$. $u_{T,k} \triangleq u_{1,k} + u_{2,k}$ denotes the total system contents at the beginning of slot k. With this definition, $\{(r_k, u_{1,k}, u_{2,k}), k \geq 1\}$ forms a Markov-chain of the *PRD* queue. A sample of the time-axis is shown in 5.1 to demonstrate the PRD priority scheduling discipline and the stochastic variables. In this example, a class-2 service time of 5 slots is preempted by a newly arriving class-1 packet and it is repeated with a new sample of 3 slots after this class-1 packet's service time.

However, in the case of the *PRI* queue, this set of stochastic variables still does not form a Markov-chain, since an interrupted class-2 service time is not resampled and thus the (complete) service time has to be kept track of. Therefore, we introduce an additional stochastic variable $t_{2,k}$ for the PRI Markov-chain, as follows: $t_{2,k}$ indicates the complete service time of the oldest class-2 packet at the beginning of slot k, if $u_{2,k} > 0$, and $t_{2,k} = 0$ if $u_{2,k} = 0$. With this

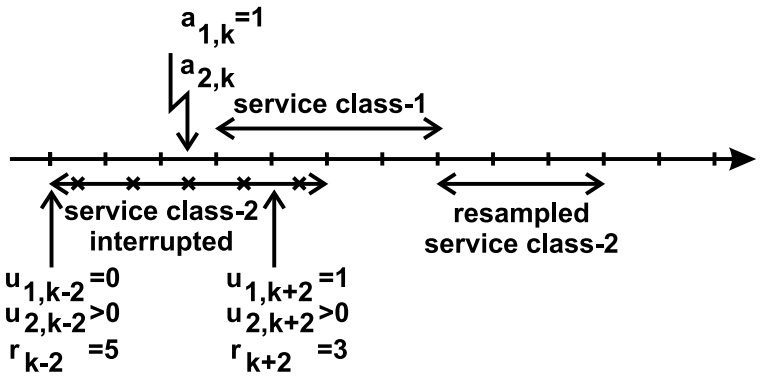


Figure 5.1: Sample of the time-axis in case of the PRD priority queue

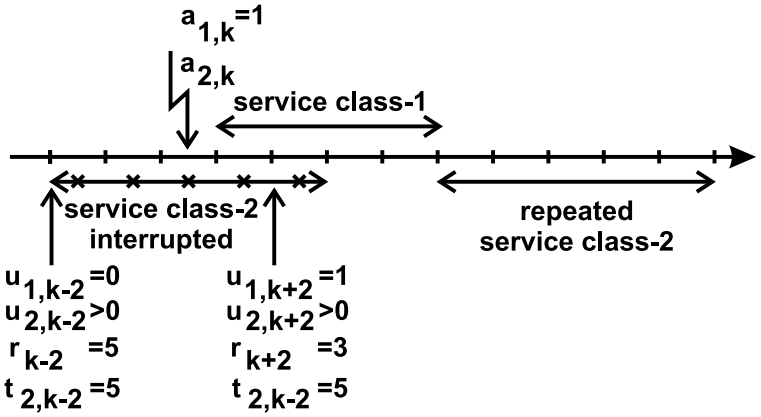


Figure 5.2: Sample of the time-axis in case of the PRI priority queue

definition, $\{(r_k, u_{1,k}, t_{2,k}, u_{2,k}), k \geq 1\}$ forms a Markov-chain of the PRI priority system. A sample of the time-axis is shown in 5.2 to demonstrate the PRI priority scheduling discipline and the stochastic variables involved. In this example, a class-2 service time of 5 slots is preempted by a newly arriving class-1 packet and it is identically repeated after this class-1 packet's service time.

We first calculate the joint pgf of the steady-state version of $\{r_k, u_{1,k}, u_{2,k}\}$ in the PRD case and calculate the joint pgf of the steady-state version of $\{r_k, u_{1,k}, t_{2,k}, u_{2,k}\}$ in the PRI priority queue next. For both models, $s^*_{j,k}$ ($j = 1, 2$) denotes the service time of the next class-j packet to receive service after slot k.

5.2.1 PRD

The following system equations are established:

1. If $r_k = 0$ (and hence $u_{T,k} = 0$):

$$u_{1,k+1} = a_{1,k} \tag{5.1}$$
$$u_{2,k+1} = a_{2,k} \tag{5.2}$$
$$r_{k+1} = \begin{cases} 0 & \text{if } a_{1,k} = a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{1,k} = 0, a_{2,k} > 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases}. \tag{5.3}$$

   The system was empty at the beginning of slot k. If no packets arrive during slot k, the system stays empty. If no class-1 packets and at least one class-2 packet arrives during slot k, a class-2 packet starts receiving service at the beginning of slot $k+1$. If at least one class-1 packet arrives during slot k, a class-1 packet enters the server.

2. If $r_k = 1$:

   (a) If $u_{1,k} = 0$ (and thus $u_{2,k} > 0$):

$$u_{1,k+1} = a_{1,k} \tag{5.4}$$
$$u_{2,k+1} = u_{2,k} - 1 + a_{2,k} \tag{5.5}$$
$$r_{k+1} = \begin{cases} 0 & \text{if } a_{1,k} = u_{2,k} - 1 + a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{1,k} = 0, u_{2,k} - 1 + a_{2,k} > 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases}. \tag{5.6}$$

   A class-2 packet left the system at the end of slot k. A new packet (of class-1, if any, otherwise of class-2) enters the server at the beginning of slot $k+1$ (if the system is non-empty at that time instant).

   (b) If $u_{1,k} > 0$:

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \tag{5.7}$$
$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{5.8}$$
$$r_{k+1} = \begin{cases} 0 & \text{if } u_{1,k} - 1 + a_{1,k} = u_{2,k} + a_{2,k} = 0 \\ s^*_{2,k} & \text{if } u_{1,k} - 1 + a_{1,k} = 0, u_{2,k} + a_{2,k} > 0 \\ s^*_{1,k} & \text{if } u_{1,k} - 1 + a_{1,k} > 0 \end{cases}. \tag{5.9}$$

   A class-1 packet left the system at the end of slot k. A new packet commences service at the beginning of slot $k + 1$.

3. If $r_k > 1$:

(a) If $u_{1,k} = 0$ (and thus $u_{2,k} > 0$):

$$u_{1,k+1} = a_{1,k} \tag{5.10}$$
$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{5.11}$$
$$r_{k+1} = \begin{cases} r_k - 1 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases}. \tag{5.12}$$

A class-2 packet was in service during slot k, but needs at least one more slot to complete its service. It stays in the server when no new class-1 packets arrive during slot k, otherwise its service is preempted by one of the newly arriving class-1 packets.

(b) If $u_{1,k} > 0$:

$$u_{1,k+1} = u_{1,k} + a_{1,k} \tag{5.13}$$
$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{5.14}$$
$$r_{k+1} = r_k - 1. \tag{5.15}$$

A class-1 packet was in service during slot k, but needs at least one more slot to complete its service. Its residual service time is decreased by one.

We define

$$P_k(x, z_1, z_2) \triangleq \mathrm{E}[x^{r_k} z_1^{u_{1,k}} z_2^{u_{2,k}}], \tag{5.16}$$

as the joint pgf of the state vector $(r_k, u_{1,k}, u_{2,k})$. We also define the partial pgf's

$$R_{1,k}(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k}-1} z_2^{u_{2,k}} \{r_k = 1, u_{1,k} > 0\}\right] \tag{5.17}$$
$$R_{2,k}(z_2) \triangleq \mathrm{E}\left[z_2^{u_{2,k}-1} \{r_k = 1, u_{1,k} = 0\}\right]. \tag{5.18}$$

Using the system equations, we constitute the following relation between $P_k(.,.,.)$ and $P_{k+1}(.,.,.)$:

$$\begin{aligned} P_{k+1}(x, z_1, z_2) =& A(0,0)P_k(0,0,0) + (A(0,z_2) - A(0,0))S_2(x)P_k(0,0,0) \\ &+ (A(z_1,z_2) - A(0,z_2))S_1(x)P_k(0,0,0) + A(0,0)R_{2,k}(0) \\ &+ [A(0,z_2)R_{2,k}(z_2) - A(0,0)R_{2,k}(0)]S_2(x) \\ &+ (A(z_1,z_2) - A(0,z_2))S_1(x)R_{2,k}(z_2) + A(0,0)R_{1,k}(0,0) \\ &+ [A(0,z_2)R_{1,k}(0,z_2) - A(0,0)R_{1,k}(0,0)]S_2(x) \\ &+ [A(z_1,z_2)R_{1,k}(z_1,z_2) - A(0,z_2)R_{1,k}(0,z_2)]S_1(x) \end{aligned} \tag{5.19}$$

$$
\begin{aligned}
&+ \frac{A(0,z_2)}{x}[P_k(x,0,z_2) - xz_2R_{2,k}(z_2) - P_k(0,0,0)] \\
&+ (A(z_1,z_2) - A(0,z_2))[P_{1,k}(1,0,z_2) - z_2R_{2,k}(z_2) \\
&- P_k(0,0,0)]S_1(x) + \frac{A(z_1,z_2)}{x}[P_k(x,z_1,z_2) \\
&- xz_1R_{1,k}(z_1,z_2) - P_k(x,0,z_2)].
\end{aligned}
$$

We assume that the system is stable (we will comment on the stability condition later) and as a result $P_k(x,z_1,z_2)$, $R_{1,k}(z_1,z_2)$ and $R_{2,k}(z_2)$ converge to steady-state functions denoted by $P(x,z_1,z_2)$, $R_1(z_1,z_2)$ and $R_2(z_2)$ respectively. By taking the $k \to \infty$ limit in expression (5.19), we obtain:

$$
\begin{aligned}
P(x,z_1,z_2) =& \frac{1}{x - A(z_1,z_2)}\Big\{xA(0,0)(1-S_2(x))[P(0,0,0) + R_1(0,0) + R_2(0)] \\
&+ A(0,z_2)(xS_2(x)-1)P(0,0,0) \\
&+ xA(0,z_2)(S_2(x) - S_1(x))R_1(0,z_2) + x[A(0,z_2)(S_2(x) - z_2) \\
&+ (A(z_1,z_2) - A(0,z_2))S_1(x)(1-z_2)]R_2(z_2) \\
&+ (A(z_1,z_2) - A(0,z_2))[xS_1(x)P(1,0,z_2) - P(x,0,z_2)] \\
&+ xA(z_1,z_2)(S_1(x) - z_1)R_1(z_1,z_2)\Big\}.
\end{aligned} \tag{5.20}
$$

It now remains for us to determine the unknown functions $P(x,0,z_2)$, $R_1(z_1,z_2)$ and $R_2(z_2)$. This can be done in the following steps. Firstly, we observe that $P(x,0,0) = P(0,0,0)$, due to the fact that $r_k = 0$ iff $u_{1,k} = u_{2,k} = 0$. By putting $z_j = 0$ ($j = 1,2$) in (5.20) and using this observation, we obtain:

$$
P(0,0,0) = A(0,0)\left[P(0,0,0) + R_1(0,0) + R_2(0)\right]. \tag{5.21}
$$

Replacing z_1 by 0 in equation (5.20) and using equation (5.21), we find the following expression for $P(x,0,z_2)$:

$$
\begin{aligned}
P(x,0,z_2) =& \frac{1}{x - A(0,z_2)}\{[x(1-S_2(x)) + A(0,z_2)(xS_2(x)-1)]P(0,0,0) \\
&+ xA(0,z_2)S_2(x)R_1(0,z_2) + xA(0,z_2)(S_2(x) - z_2)R_2(z_2)\}.
\end{aligned} \tag{5.22}
$$

We note that $P(x,0,z_2)$ is bound for all values of x and z_2 such that $|x| < 1$ and $|z_2| < 1$ since $P(x,z_1,z_2)$ is a pgf. In particular, this should be true for $x = A(0,z_2)$, $|z_2| < 1$, since $|A(0,z_2)| < 1$ for all such z_2. If we choose $x = A(0,z_2)$ in equation (5.22), with $|z_2| < 1$, the denominator in the right-hand side of this equation equals zero. The above then implies that its numerator also equals

zero, which yields the following expression:

$$R_1(0,z_2) = \frac{E_2(0,z_2)(1-A(0,z_2))P(0,0,0) + A(0,z_2)(z_2 - E_2(0,z_2))R_2(z_2)}{A(0,z_2)E_2(0,z_2)}, \tag{5.23}$$

with $E_j(z_1,z_2) \triangleq S_j(A(z_1,z_2))$.

Returning to expression (5.20), we notice that $P(x,z_1,z_2)$ must be bound for all values of x and z_j such that $|x|<1$ and $|z_j|<1$ ($j=1,2$) since $P(x,z_1,z_2)$ is a pgf. In particular, this should be true for $x = A(z_1,z_2)$, $|z_j|<1$ ($j=1,2$), since $|A(z_1,z_2)|<1$ for all such z_j. Substituting x by $A(z_1,z_2)$ in equation (5.20), where $|z_j|<1$, the denominator in the right-hand side of this equation vanishes. The same must then be true for its numerator, which yields the following expression:

$$R_1(z_1,z_2) = \frac{E_1(z_1,z_2)\left\{\begin{array}{l}(A(0,z_2)-1)(A(z_1,z_2)-1)E_2(0,z_2)P(0,0,0)\\ +[(A(z_1,z_2)-A(0,z_2))E_2(0,z_2)(z_2-1)\\ -A(0,z_2)(A(z_1,z_2)-1)(z_2-E_2(0,z_2))]R_2(z_2)\end{array}\right\}}{(A(0,z_2)-1)A(z_1,z_2)E_2(0,z_2)(z_1-E_1(z_1,z_2))}, \tag{5.24}$$

by substituting $P(x,0,z_2)$ and $R_1(0,z_2)$ by their expressions (5.22) and (5.23) respectively. We furthermore note that $R_1(z_1,z_2)$ must be bound for all values of z_j such that $|z_j|<1$ ($j=1,2$). In particular, this should be true for $z_1 = Y_1(z_2)$ - with

$$Y_1(z) \triangleq E_1(Y_1(z),z), \tag{5.25}$$

- and $|z_2|<1$ (see the Appendix for more details). The above implies that if we insert $z_1 = Y_1(z_2)$ in equation (5.24), where $|z_2|<1$, the denominator in the right-hand side of this equation vanishes. The same must then be true for its numerator, yielding

$$R_2(z_2) = P(0,0,0)\frac{(A(Y_1(z_2),z_2)-1)Y_2(z_2)}{A(Y_1(z_2),z_2)(z_2-Y_2(z_2))}, \tag{5.26}$$

with

$$Y_2(z) = \frac{(1-A(0,z))A(Y_1(z),z)S_2(A(0,z))}{(A(Y_1(z),z)-A(0,z))S_2(A(0,z)) - A(0,z)(A(Y_1(z),z)-1)}. \tag{5.27}$$

The following expression of $P(x, z_1, z_2)$ is derived by substituting the equations (5.21)-(5.26) in expression (5.20):

$$\begin{aligned} P(x,z_1,z_2) = P(0,0,0)\Bigg\{1 \\ + xz_1\frac{(A(z_1,z_2)-A(Y_1(z_2),z_2))(S_1(x)-E_1(z_1,z_2))Y_2(z_2)(z_2-1)}{A(Y_1(z_2),z_2)(x-A(z_1,z_2))(z_1-E_1(z_1,z_2))(z_2-Y_2(z_2))} \\ +xz_2\frac{A(0,z_2)(A(Y_1(z_2),z_2)-1)(S_2(x)-E_2(0,z_2))Y_2(z_2)}{A(Y_1(z_2),z_2)E_2(0,z_2)(x-A(0,z_2))(z_2-Y_2(z_2))}\Bigg\}. \end{aligned} \tag{5.28}$$

Finally, in order to find an expression for $P(0,0,0)$, we put $x = z_1 = z_2 = 1$ in equation (5.28) and use de l'Hôpital's rule. We obtain

$$P(0,0,0) = 1 - \rho_{T,eff}, \tag{5.29}$$

with

$$\rho_{T,eff} \triangleq \rho_1 + \lambda_2\mu_{2,eff}, \tag{5.30}$$

and

$$\mu_{2,eff} \triangleq \frac{A_1(0)(1-S_2(A_1(0)))}{S_2(A_1(0))(1-A_1(0))}. \tag{5.31}$$

So, $\rho_{T,eff}$ is the effective load (including retransmissions of class-2 packets) offered to the system. Using this result in equation (5.28), we finally obtain an expression of $P(x, z_1, z_2)$ in terms of the system parameters.

5.2.2 PRI

As already mentioned, we need an extra stochastic variable $t_{2,k}$ - the complete service time of the oldest class-2 packet - in order to analyze the PRI priority queue.

The following system equations are established:

1. If $r_k = 0$ (and hence $u_{T,k} = 0$):

$$u_{1,k+1} = a_{1,k} \tag{5.32}$$

$$u_{2,k+1} = a_{2,k} \tag{5.33}$$

$$r_{k+1} = \begin{cases} 0 & \text{if } a_{1,k} = a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{1,k} = 0, a_{2,k} > 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases} \tag{5.34}$$

$$t_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases}. \tag{5.35}$$

The system is empty at the beginning of slot k. If no packets arrive during slot k, the system stays empty. If no class-1 packets and at least one class-2 packet arrive during slot k, a class-2 packet starts receiving service at the beginning of slot $k+1$. If at least one class-1 packet arrives during slot k, a class-1 packet enters the server. If class-2 packets arrive, $t_{2,k+1}$ equals the service time of one of these class-2 packets.

2. If $r_k = 1$:

   (a) If $u_{1,k} = 0$ (and thus $u_{2,k} > 0$):

$$u_{1,k+1} = a_{1,k} \tag{5.36}$$

$$u_{2,k+1} = u_{2,k} - 1 + a_{2,k} \tag{5.37}$$

$$r_{k+1} = \begin{cases} 0 & \text{if } a_{1,k} = u_{2,k} - 1 + a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{1,k} = 0, u_{2,k} - 1 + a_{2,k} > 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases} \tag{5.38}$$

$$t_{2,k+1} = \begin{cases} 0 & \text{if } u_{2,k} - 1 + a_{2,k} = 0 \\ s^*_{2,k} & \text{if } u_{2,k} - 1 + a_{2,k} > 0 \end{cases}. \tag{5.39}$$

   A class-2 packet is served during slot k and leaves the system at the end of this slot. A new packet (of class-1, if any, otherwise of class-2) enters the server at the beginning of slot $k+1$ (if the system is non-empty at that time instant).

   (b) If $u_{2,k} = 0$ (and thus $u_{1,k} > 0$):

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \tag{5.40}$$

$$u_{2,k+1} = a_{2,k} \tag{5.41}$$

$$r_{k+1} = \begin{cases} 0 & \text{if } u_{1,k} - 1 + a_{1,k} = a_{2,k} = 0 \\ s^*_{2,k} & \text{if } u_{1,k} - 1 + a_{1,k} = 0, a_{2,k} > 0 \\ s^*_{1,k} & \text{if } u_{1,k} - 1 + a_{1,k} > 0 \end{cases} \tag{5.42}$$

$$t_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases}. \tag{5.43}$$

   A class-1 packet leaves the system at the end of slot k. A new packet commences service at the beginning of slot $k+1$.

   (c) If $u_{1,k} > 0, u_{2,k} > 0$:

$$u_{1,k+1} = u_{1,k} - 1 + a_{1,k} \tag{5.44}$$

$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{5.45}$$

$$r_{k+1} = \begin{cases} t_{2,k} \text{ if } u_{1,k} - 1 + a_{1,k} = 0 \\ s^*_{1,k} \text{ if } u_{1,k} - 1 + a_{1,k} > 0 \end{cases} \tag{5.46}$$

$$t_{2,k+1} = t_{2,k}. \tag{5.47}$$

The difference with the previous situation is that at least one class-2 packet is in the system at the beginning of slot k. The oldest class-2 packet - with service time equal to $t_{2,k}$ - starts service at the beginning of the next slot if no class-1 packets are present in the system at that time instant.

3. If $r_k > 1$:

   (a) If $u_{1,k} = 0$ (and thus $u_{2,k} > 0$):

$$u_{1,k+1} = a_{1,k} \tag{5.48}$$

$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{5.49}$$

$$r_{k+1} = \begin{cases} r_k - 1 & \text{if } a_{1,k} = 0 \\ s^*_{1,k} & \text{if } a_{1,k} > 0 \end{cases} \tag{5.50}$$

$$t_{2,k+1} = t_{2,k}. \tag{5.51}$$

   A class-2 packet is in service during slot k, but needs at least one more slot to complete its service. The class-2 packet stays in the server when no new class-1 packets arrive during slot k, otherwise its service is preempted by one of the newly arriving class-1 packets and the class-2 packet is put back in the queue.

   (b) If $u_{2,k} = 0$ (and thus $u_{1,k} > 0$):

$$u_{1,k+1} = u_{1,k} + a_{1,k} \tag{5.52}$$

$$u_{2,k+1} = a_{2,k} \tag{5.53}$$

$$r_{k+1} = r_k - 1 \tag{5.54}$$

$$t_{2,k+1} = \begin{cases} 0 & \text{if } a_{2,k} = 0 \\ s^*_{2,k} & \text{if } a_{2,k} > 0 \end{cases}. \tag{5.55}$$

   A class-1 packet is in service during slot k, but needs at least one more slot to complete its service.

   (c) If $u_{1,k} > 0, u_{2,k} > 0$:

$$u_{1,k+1} = u_{1,k} + a_{1,k} \tag{5.56}$$

$$u_{2,k+1} = u_{2,k} + a_{2,k} \tag{5.57}$$

$$r_{k+1} = r_k - 1 \tag{5.58}$$

$$t_{2,k+1} = t_{2,k}. \tag{5.59}$$

The difference with the previous situation is that at least one class-2 packet was present in the system at the beginning of slot k.

We define

$$P_k(x, z_1, y_2, z_2) \triangleq \mathrm{E}[x^{r_k} z_1^{u_{1,k}} y_2^{t_{2,k}} z_2^{u_{2,k}}], \tag{5.60}$$

as the joint pgf of the state vector $(r_k, u_{1,k}, t_{2,k}, u_{2,k})$. Further, we define the partial pgf's

$$R_{1,k}(z_1, y_2, z_2) \triangleq \mathrm{E}\left[z_1^{u_{1,k}-1} y_2^{t_{2,k}} z_2^{u_{2,k}} \{r_k = 1, u_{1,k} > 0\}\right] \tag{5.61}$$

$$R_{2,k}(y_2, z_2) \triangleq \mathrm{E}\left[y_2^{t_{2,k}} z_2^{u_{2,k}-1} \{r_k = 1, u_{1,k} = 0\}\right]. \tag{5.62}$$

Using the system equations, we constitute the following relation between $P_k(x, z_1, y_2, z_2)$ and $P_{k+1}(x, z_1, y_2, z_2)$:

$$\begin{aligned}
P_{k+1}(x,z_1, y_2, z_2) &= [A(0,0) + (A(0,z_2) - A(0,0))S_2(xy_2) \\
&+ (A(z_1,0) - A(0,0))S_1(x) + (A(z_1,z_2) - A(z_1,0) - A(0,z_2) \\
&+ A(0,0))S_1(x)S_2(y_2)]P_k(0,0,0,0) + A(0,0)R_{2,k}(1,0) \\
&+ [A(0,z_2)R_{2,k}(1,z_2) - A(0,0)R_{2,k}(1,0)]S_2(xy_2) \\
&+ (A(z_1,0) - A(0,0))S_1(x)R_{2,k}(1,0) + [(A(z_1,z_2) \\
&- A(0,z_2))R_{2,k}(1,z_2) - (A(z_1,0) - A(0,0))R_{2,k}(1,0)]S_1(x)S_2(y_2) \\
&+ [A(0,0) + (A(0,z_2) - A(0,0))S_2(xy_2)]R_{1,k}(0,0,0) \\
&+ [A(z_1,0)R_{1,k}(z_1,0,0) - A(0,0)R_{1,k}(0,0,0)]S_1(x) + [(A(z_1,z_2) \\
&- A(z_1,0))R_{1,k}(z_1,0,0) - (A(0,z_2) - A(0,0))R_{1,k}(0,0,0)]S_1(x)S_2(y_2) \\
&+ A(0,z_2)(R_{1,k}(0,xy_2,z_2) - R_{1,k}(0,0,0)) + [A(z_1,z_2)(R_{1,k}(z_1,y_2,z_2) \\
&- R_{1,k}(z_1,0,0)) - A(0,z_2)(R_{1,k}(0,y_2,z_2) - R_{1,k}(0,0,0))]S_1(x) \\
&+ \frac{A(0,z_2)}{x}[P_k(x,0,y_2,z_2) - xz_2R_{2,k}(y_2,z_2) - P_k(0,0,0,0)] \\
&+ (A(z_1,z_2) - A(0,z_2))S_1(x)[P_{1,k}(1,0,y_2,z_2) - z_2R_{2,k}(y_2,z_2) \\
&- P_k(0,0,0,0)] + \frac{A(z_1,0) + (A(z_1,z_2) - A(z_1,0))S_2(y_2)}{x} \\
&\times [P_k(x,z_1,0,0) - xz_1R_{1,k}(z_1,0,0) - P_k(0,0,0,0)] \\
&+ \frac{A(z_1,z_2)}{x}[(P_k(x,z_1,y_2,z_2) - P_k(x,z_1,0,0)) - xz_1(R_{1,k}(z_1,y_2,z_2) \\
&- R_1(z_1,0,0)) - (P_k(x,0,y_2,z_2) - P_k(0,0,0,0))].
\end{aligned} \tag{5.63}$$

We assume that the system is stable (we will comment on the stability condition in subsection 5.2.3) and as a result $P_k(x, z_1, y_2, z_2)$, $R_{1,k}(z_1, y_2, z_2)$ and $R_{2,k}(y_2, z_2)$ converge to the steady-state functions $P(x, z_1, y_2, z_2)$, $R_1(z_1, y_2, z_2)$

and $R_2(y_2, z_2)$ respectively. By taking the $k \to \infty$ limit in expression (5.63), we obtain:

$$\begin{aligned} P(x, z_1, y_2, z_2) =& \frac{1}{x - A(z_1, z_2)} \Big\{ xA(0,0)(1 - S_2(xy_2) + S_1(x)(S_2(y_2) - 1)) \\ & \times (P(0,0,0,0) + R_1(0,0,0) + R_2(1,0)) \\ & + [A(0, z_2)(xS_2(xy_2) - 1 + xS_1(x)(1 - S_2(y_2))) \\ & + (A(z_1, z_2) - A(z_1, 0))(xS_1(x) - 1)(S_2(y_2) - 1)]P(0,0,0,0) \\ & + xA(z_1, 0)S_1(x)(1 - S_2(y_2))R_2(1,0) + xA(0, z_2)(S_2(xy_2) - 1 \\ & + S_1(x)(1 - S_2(y_2)))R_1(0,0,0) + x(A(z_1, z_2) - A(z_1, 0)) \\ & \times (S_1(x) - z_1)(S_2(y_2) - 1)R_1(z_1, 0, 0) + (A(z_1, z_2) - A(z_1, 0)) \\ & \times (S_2(y_2) - 1)P(x, z_1, 0, 0) + xA(0, z_2)S_2(xy_2)R_2(1, z_2) \\ & + x(A(z_1, z_2) - A(0, z_2))S_1(x)S_2(y_2)R_2(1, z_2) \\ & - xA(0, z_2)S_1(x)R_1(0, y_2, z_2) - xz_2[A(0, z_2) \\ & + (A(z_1, z_2) - A(0, z_2))S_1(x)]R_2(y_2, z_2) + x(A(z_1, z_2) \\ & - A(0, z_2))S_1(x)P(1, 0, y_2, z_2) - (A(z_1, z_2 - A(0, z_2)) \\ & \times P(x, 0, y_2, z_2) + xA(0, z_2)R_1(0, xy_2, z_2) \\ & + xA(z_1, z_2)(S_1(x) - z_1)R_1(z_1, y_2, z_2) \Big\}. \end{aligned} \tag{5.64}$$

In the remainder we determine the functions $P(x, z_1, 0, 0)$, $P(x, 0, y_2, z_2)$, $R_1(z_1, y_2, z_2)$ and $R_2(y_2, z_2)$ and the unknown constants $P(0,0,0,0)$, $R_1(0,0,0)$ and $R_2(1,0)$. This can be done in a few steps. Firstly, we observe that $P(x, 0, y_2, 0) = P(0,0,0,0)$ and $R_1(0, y_2, 0) = R_1(0,0,0)$, due to the fact that $r_k = 0$ iff $u_{1,k} = u_{2,k} = 0$ and $t_{2,k} = 0$ iff $u_{2,k} = 0$. By putting $z_j = 0$ ($j = 1, 2$) in expression (5.64) and using this observation, we obtain:

$$P(0,0,0,0) = A(0,0)\left[P(0,0,0,0) + R_1(0,0,0) + R_2(1,0)\right]. \tag{5.65}$$

We furthermore observe that the following equations hold (because $t_{2,k} = 0$ iff $u_{2,k} = 0$):

$$P(x, z_1, y_2, 0) = P(x, z_1, 0, 0) \tag{5.66}$$

$$R_1(z_1, y_2, 0) = R_1(z_1, 0, 0). \tag{5.67}$$

Replacing z_2 (z_1 respectively) by 0 in equation (5.64) and using the former equations and equation (5.65), we find the following expression for $P(x, z_1, 0, 0)$

($P(x,0,y_2,z_2)$ respectively):

$$P(x,z_1,0,0) = \frac{\left\{\begin{array}{l}\big[x(1-S_1(x)) + A(z_1,0)(xS_1(x)-1)\big]P(0,0,0,0) \\ +xA(z_1,0)\big[S_1(x)R_2(1,0) + (S_1(x)-z_1)R_1(z_1,0,0)\big]\end{array}\right\}}{x - A(z_1,0)} \tag{5.68}$$

$$P(x,0,y_2,z_2) = \frac{\left\{\begin{array}{l}\big[x(1-S_2(xy_2)) + A(0,z_2)(xS_2(xy_2)-1)\big]P(0,0,0,0) \\ +xA(0,z_2)\big[S_2(xy_2)R_2(1,z_2) - z_2R_2(y_2,z_2) \\ +(S_2(xy_2)-1)R_1(0,0,0) + R_1(0,xy_2,z_2)\big]\end{array}\right\}}{x - A(0,z_2)}. \tag{5.69}$$

Substituting these expressions in (5.64) allows us to eliminate $P(x,z_1,0,0)$, $P(x,0,y_2,z_2)$ and $P(1,0,y_2,z_2)$.

We now first return to expression (5.64) and extract as much information as possible from this expression. We notice that the function $P(x,z_1,y_2,z_2)$ must be bound for all values of x, y_2 and z_j such that $|x|<1$, $|y_2|<1$ and $|z_j|<1$ ($j=1,2$) since $P(x,z_1,y_2,z_2)$ is a pgf. In particular, this should be true for $x = A(z_1,z_2)$, $|x|<1$, $|y_2|<1$ and $|z_j|<1$ ($j=1,2$), since $|A(z_1,z_2)|<1$ for $|z_j|<1$. The above implies that if we choose $x = A(z_1,z_2)$ in equation (5.64) the denominator of the right-hand side of this equation vanishes. The same is then true for the numerator, which yields the following expression (by also substituting $P(x,z_1,0,0)$ and $P(x,0,y_2,z_2)$ by their expressions obtained in equations (5.68) and (5.69) respectively):

$$\begin{aligned}(S_2(y_2)-1)R_1(z_1,0,0) + R_1(z_1,y_2,z_2) \\ = \frac{E_1(z_1,z_2)}{A(z_1,z_2)(1-A(0,z_2))(z_1-E_1(z_1,z_2))} \\ \times \{A(0,z_2)(1-A(z_1,z_2))[(1-S_2(y_2))R_1(0,0,0) - R_1(0,y_2,z_2)] \\ + (A(z_1,z_2)-A(0,z_2))[S_2(y_2)R_2(1,z_2) - z_2R_2(y_2,z_2)]\}.\end{aligned} \tag{5.70}$$

Next, we notice that $(S_2(y_2)-1)R_1(z_1,0,0) + R_1(z_1,y_2,z_2)$ must be bound for all values of y_2 and z_j such that $|y_2|<1$ and $|z_j|<1$ ($j=1,2$). In particular, this is true for $z_1 = Y_1(z_2)$, with

$$Y_1(z) \triangleq E_1(Y_1(z),z). \tag{5.71}$$

The above implies that if we insert $z_1 = Y_1(z_2)$ in equation (5.70), where $|z_2|<1$, the denominator of the right-hand side of this equation equals zero. The same must then be true its numerator, yielding

$$(S_2(y_2)-1)R_1(0,0,0) + R_1(0,y_2,z_2) \tag{5.72}$$

$$= \frac{(A(Y_1(z_2), z_2) - A(0, z_2))(S_2(y_2)R_2(1, z_2) - z_2 R_2(y_2, z_2))}{A(0, z_2)(1 - A(Y_1(z_2), z_2))}.$$

From this expression, no more information can be extracted. Therefore, we return to expressions (5.68) and (5.69).

Firstly, we notice that expression (5.68) must be bound for all values of x and z_1 such that $|x| < 1$ and $|z_1| < 1$. In particular, this should be true for $x = A(z_1, 0)$. The above implies that if we choose $x = A(z_1, 0)$ in equation (5.68), where $|z_1| < 1$, the denominator of the right-hand side of this equation vanishes. Of course, the same must then be true for its numerator, which yields the following relation for $R_1(z_1, 0, 0)$:

$$R_1(z_1, 0, 0) = \frac{E_1(z_1, 0)[(A(z_1, 0) - 1)P(0, 0, 0, 0) + A(z_1, 0)R_2(1, 0)]}{A(z_1, 0)(z_1 - E_1(z_1, 0))}. \quad (5.73)$$

This expression must be bound for its argument z_1 within the complex unit circle, the denominator of the right-hand side of this expression vanishes when $z_1 = Y_1(0)$, since the denominator vanishes also for this value of z_1 - since $Y(0) = E_1(Y(0), 0)$ - yielding:

$$R_2(1, 0) = P(0, 0, 0, 0)\frac{1 - A(Y_1(0), 0)}{A(Y_1(0), 0)}. \quad (5.74)$$

We then determine the following expression for $P(x, z_1, 0, 0)$ from equation (5.68) together with equations (5.73) and (5.74):

$$\begin{aligned} P(x, z_1, 0, 0) =& P(0, 0, 0, 0) \quad (5.75) \\ & \times \left[1 + xz_1 \frac{(A(z_1, 0) - A(Y_1(0), 0))(S_1(x) - S_1(A(z_1), 0))}{A(Y_1(0), 0)(x - A(z_1, 0))(z_1 - E_1(z_1, 0))}\right]. \end{aligned}$$

Finally, we find the remaining unknown function $R_2(y_2, z_2)$ (which is the only still unknown - except for $P(0, 0, 0, 0)$ - at this time). First, substituting expression (5.72) in expression (5.69) yields the following expression of $P(x, 0, y_2, z_2)$:

$$P(x, 0, y_2, z_2) = \frac{\left\{\begin{array}{l} (x - A(0, z_2))(1 - A(Y_1(z_2), z_2))P(0, 0, 0, 0) \\ -x(1 - A(0, z_2))S_2(xy_2)(1 - A(Y_1(z_2), z_2))P(0, 0, 0, 0) \\ +x(1 - A(0, z_2))A(Y_1(z_2), z_2)S_2(xy_2)R_2(1, z_2) \\ -xz_2A(0, z_2)(1 - A(Y_1(z_2), z_2))R_2(y_2, z_2) \\ -xz_2(A(Y_1(z_2), z_2) - A(0, z_2)R_2(xy_2, z_2) \end{array}\right\}}{(x - A(0, z_2))(1 - A(Y_1(z_2), z_2))}. \quad (5.76)$$

$P(x, 0, y_2, z_2)$ must be bound for all values of x, y_2 and z_2 such that $|x| < 1$, $|y_2| < 1$ and $|z_2| < 1$. In particular, this should be true for $x = A(0, z_2)$.

Choosing $x = A(0, z_2)$ in equation (5.76), the denominator of the right-hand side vanishes. The same must be true for the denominator, yielding:

$$R_2(A(0,z_2)y_2,z_2) = \frac{\left\{\begin{array}{l}(1-A(0,z_2))S_2(A(0,z_2)y_2)[(A(Y_1(z_2),z_2)-1)\\ \times P(0,0,0,0)+A(Y_1(z_2),z_2)R_2(1,z_2)]\\ +A(0,z_2)z_2(A(Y_1(z_2),z_2)-1)R_2(y_2,z_2)\end{array}\right\}}{z_2(A(Y_1(z_2),z_2)-A(0,z_2))}. \tag{5.77}$$

This latter expression thus gives a functional equation of $R_2(y_2, z_2)$ (as a function of $P(0, 0, 0, 0)$). Defining the following partial conditional pgf

$$R_{2,i}(z_2) \triangleq \mathrm{E}\left[z_2^{u_2-1}|t_2 = i\{r = 1, u_1 = 0\}\right], \tag{5.78}$$

where r, u_1, t_2 and u_2 are the steady state versions of r_k, $u_{1,k}$, $t_{2,k}$ and $u_{2,k}$ respectively. $R_2(y_2, z_2)$ is expressed in function of the $R_{2,i}(z_2)$ as follows:

$$R_2(y_2,z_2) = \sum_{i=1}^{\infty} \mathrm{Prob}[t_2 = i|r = 1, u_1 = 0]y_2^i R_{2,i}(z_2). \tag{5.79}$$

Note that the slots at the beginning of which $r = 1$ and $u_1 = 0$ are the slots of which at the end a class-2 packet leaves the system (and vice versa). Since t_2 equals the service time of the packet that leaves the system (the oldest one) and since every packet leaves the system just once, we find

$$\mathrm{Prob}[t_2 = i|r = 1, u_1 = 0] = s_2(i), \tag{5.80}$$

with $s_2(i)$ the pmf of the class-2 service times. Expression (5.79) thus becomes

$$R_2(y_2,z_2) = \sum_{i=1}^{\infty} s_2(i)y_2^i R_{2,i}(z_2). \tag{5.81}$$

Substituting this expression in expression (5.77) gives

$$\begin{aligned} z_2(A(Y_1(z_2),z_2)-A(0,z_2))&\sum_{i=1}^{\infty} s_2(i)(A(0,z_2)y_2)^i R_{2,i}(z_2) \qquad (5.82)\\ =&(1-A(0,z_2))\sum_{i=1}^{\infty} s_2(i)(A(0,z_2)y_2)^i\\ &\times[(A(Y_1(z_2),z_2)-1)P(0,0,0,0)+A(Y_1(z_2),z_2)R_2(1,z_2)]\\ &+A(0,z_2)z_2(A(Y_1(z_2),z_2)-1)\sum_{i=1}^{\infty} s_2(i)y_2^i R_{2,i}(z_2), \end{aligned}$$

where we have also substituted $S_2(A(0,z_2)y_2)$ by its power series expression. Since this equation has to hold for all y_2 ($|y_2|<1$), the coefficients of y_2^i in both sides of the expression have to be equal (and this for all i). This leads to

$$R_{2,i}(z_2) = \frac{(1-A(0,z_2))A(0,z_2)^i \left\{ \begin{array}{l} (A(Y_1(z_2),z_2)-1)P(0,0,0,0) \\ +A(Y_1(z_2),z_2)R_2(1,z_2) \end{array} \right\}}{z_2[(A(Y_1(z_2),z_2)-A(0,z_2))A(0,z_2)^i - A(0,z_2)(A(Y_1(z_2),z_2)-1)]}, \tag{5.83}$$

$i \geq 1$. Substituting this in (5.81) yields

$$R_2(y_2,z_2) = \frac{B(y_2,z_2)}{z_2}\left(\frac{A(Y_1(z_2),z_2)-1}{A(Y_1(z_2),z_2)}P(0,0,0,0) + R_2(1,z_2)\right), \tag{5.84}$$

with

$$B(y,z) \triangleq \sum_{i=1}^{\infty} \frac{s_2(i)y^i(1-A(0,z))A(Y_1(z),z)A(0,z)^i}{(A(Y_1(z),z)-A(0,z))A(0,z)^i - A(0,z)(A(Y_1(z),z)-1)}. \tag{5.85}$$

Substituting y_2 by 1 in expression (5.84), gives an expression of $R_2(1,z_2)$ as a function of $P(0,0,0,0)$:

$$R_2(1,z_2) = P(0,0,0,0)\frac{(A(Y_1(z_2),z_2)-1)B(1,z_2)}{A(Y_1(z_2),z_2)(z_2-B(1,z_2))}. \tag{5.86}$$

Substituting expressions (5.84) and (5.86) in expression (5.76) yields the following expression of $P(x,0,y_2,z_2)$ as a function of $P(0,0,0,0)$:

$$\begin{aligned} &P(x,0,y_2,z_2) \qquad (5.87)\\ &= P(0,0,0,0)\left[1 + xz_2\frac{\left\{ \begin{array}{l} A(Y_1(z_2),z_2)(B(xy_2,z_2)-S_2(xy_2)) \\ +A(0,z_2)A(Y_1(z_2),z_2)(S_2(xy_2)-B(y_2,z_2)) \\ +A(0,z_2)(B(y_2,z_2)-B(xy_2,z_2)) \end{array} \right\}}{A(Y_1(z_2),z_2)(x-A(0,z_2))(z_2-B(1,z_2))}\right]. \end{aligned}$$

We find the following expression for $P(x,z_1,y_2,z_2)$ - since all unknown functions and constants in expression (5.64) are basically found as a function of $P(0,0,0,0)$ -

$$P(x,z_1,y_2,z_2) = P(0,0,0,0)\left[1 \right. \tag{5.88}$$

$$
\begin{aligned}
&+ xz_1 \frac{(A(z_1,0) - A(Y_1(0),0))(E_1(z_1,0) - S_1(x))(S_2(y_2) - 1)}{A(Y_1(0),0)(x - A(z_1,0))(z_1 - E_1(z_1,0))} \\
&+ xz_1 \frac{(A(z_1,z_2) - A(Y_1(z_2),z_2))(S_1(x) - E_1(z_1,z_2))}{A(Y_1(z_2),z_2)(x - A(z_1,z_2))(z_1 - E_1(z_1,z_2))(z_2 - B(1,z_2))} \\
&\times (z_2 B(y_2,z_2) - S_2(y_2)B(1,z_2)) \\
&+ xz_2 \frac{\left\{ \begin{array}{l} A(Y_1(z_2),z_2)(B(xy_2,z_2) - S_2(xy_2)) \\ +A(0,z_2)A(Y_1(z_2),z_2)(S_2(xy_2) - B(y_2,z_2)) \\ +A(0,z_2)(B(y_2,z_2) - B(xy_2,z_2)) \end{array} \right\}}{A(Y_1(z_2),z_2)(x - A(0,z_2))(z_2 - B(1,z_2))} \Bigg].
\end{aligned}
$$

Finally, in order to find an expression for $P(0,0,0,0)$, we put $x_1 = z_1 = x_2 = z_2 = 1$ and use de l'Hôpital's rule in equation (5.88). We obtain

$$P(0,0,0,0) = 1 - \rho_{T,eff}, \tag{5.89}$$

with

$$\rho_{T,eff} \triangleq \rho_1 + \lambda_2 \mu_{2,eff}, \tag{5.90}$$

and

$$\mu_{2,eff} \triangleq \frac{S_2(1/A_1(0)) - 1}{1/A_1(0) - 1}. \tag{5.91}$$

$\rho_{T,eff}$ denotes also in this model the effective load (including retransmissions of class-2 packets) offered to the system. Using this result in equation (5.88), we finally obtain $P(x, z_1, y_2, z_2)$.

5.2.3 Stability issues

We briefly touch upon stability issues of these priority queues. We notice that $P(0,0,0) = 0$ and $P(0,0,0,0) = 0$ for $\rho_{T,eff} = 1$ in the PRD and PRI case respectively. As a result the system becomes instable for $\rho_{T,eff} \geq 1$. Note that in the PRI case

$$\mu_{2,eff} \geq \mu_2. \tag{5.92}$$

This is proved by writing expression (5.91) in terms of power series, yielding

$$\mu_{2,eff} = \sum_{n=1}^{\infty} s_2(n) \frac{(1/A_1(0))^n - 1}{1/A_1(0) - 1}, \tag{5.93}$$

and by noting that

$$\frac{(1/A_1(0))^n - 1}{1/A_1(0) - 1} \geq n, \tag{5.94}$$

for all $n \geq 1$, since $A_1(0) \leq 1$. As a result the effective load is always larger than the arrival load in a PRI priority queue. This is also intuitively clear, since retransmissions increase the total (effective) load of a system. Note that this is not necessarily the case for the PRD priority queue, since the resampling of a service time can actually lead to a decrease of the (effective) load.

5.3 System contents

5.3.1 Calculation of the pgf $U(z_1, z_2)$

In this paragraph, we calculate the joint pgf of the system contents of class-1 and class-2 packets. It is given by $P(1, z_1, z_2)$ in case of PRD and by $P(1, z_1, 1, z_2)$ in the PRI case. Both lead to the following expression for $U(z_1, z_2)$:

$$U(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{u_1} z_2^{u_2}\right] \tag{5.95}$$

$$= (1 - \rho_{T,eff}) \frac{Y_2(z_2)(z_2 - 1)}{z_2 - Y_2(z_2)} \tag{5.96}$$

$$\times \left\{ 1 + z_1 \frac{(A(z_1, z_2) - A(Y_1(z_2), z_2))(E_1(z_1, z_2) - 1)}{A(Y_1(z_2), z_2)(A(z_1, z_2) - 1)(z_1 - E_1(z_1, z_2))} \right\},$$

with $Y_2(z)$ given by

$$Y_2(z) = \frac{(1 - A(0, z))A(Y_1(z), z)S_2(A(0, z))}{(A(Y_1(z), z) - A(0, z))S_2(A(0, z)) - A(0, z)(A(Y_1(z), z) - 1)}, \tag{5.97}$$

in the PRD case and by

$$Y_2(z) \triangleq B(1, z) \tag{5.98}$$

$$= \sum_{i=1}^{\infty} \frac{s_2(i)(1 - A(0, z))A(Y_1(z), z)A(0, z)^i}{(A(Y_1(z), z) - A(0, z))A(0, z)^i - A(0, z)(A(Y_1(z), z) - 1)}, \tag{5.99}$$

in the PRI case.

We give a stochastic interpretation of this function $Y_2(z)$ - which is a pgf - further in this chapter.

Special case: geometric class-2 service times

In the special case of geometric class-2 service times, i.e.,

$$S_2(z) = \frac{(1-\beta_2)z}{1-\beta_2 z}, \tag{5.100}$$

expression (5.96) equals

$$\begin{aligned} U(z_1,z_2) = &\frac{(1-\rho_T)(1-\beta_2)(z_2-1)}{z_2 - A(Y_1(z_2),z_2)(1-\beta_2+\beta_2 z_2)}\Bigg[A(Y_1(z_2),z_2) \\ &-z_1\frac{(A(Y_1(z_2),z_2)-A(z_1,z_2))(E_1(z_1,z_2)-1)}{(A(z_1,z_2)-1)(z_1-E_1(z_1,z_2))}\Bigg], \end{aligned} \tag{5.101}$$

in the case of PRD. This is the same pgf for $U(z_1,z_2)$ as in chapter 4 for a preemptive *resume* priority queue with the same arrival and service processes. The difference between a PR and a PRD scheduling discipline is that from an interrupted low-priority packet only the not-yet-served part has to be served after the interruption in the PR priority queue, while in the PRD priority queue an interrupted class-2 service has to be repeated with a new sample of the class-2 service times. Since the geometric distribution is memoryless however, a residual service time and a new sample of the complete service time have the same (geometric) distribution, and thus the system contents in a buffer with PR priority on the one hand and PRD priority on the other hand are identically distributed. Note that the expression of $U(z_1,z_2)$ in the PRI queue is different, because in this case, a second attempt of a class-2 service time is not resampled and as a result class-2 service times are no longer memoryless after an interruption.

Special case: deterministic class-2 service times

In the special case of deterministic class-2 service times, i.e.,

$$S_2(z) = z^{\mu_2}, \tag{5.102}$$

the expressions of $Y_2(z)$ in the PRD and PRI queue are equal:

$$Y_2(z) = \frac{(1-A(0,z))A(Y_1(z),z)A(0,z)^{\mu_2}}{(A(Y_1(z),z)-A(0,z))A(0,z)^{\mu_2} - A(0,z)(A(Y_1(z),z)-1)}. \tag{5.103}$$

As a result the expressions of $U(z_1,z_2)$ in both cases are also equal. Indeed, when the class-2 service times are deterministic, resampling the class-2 service times after an interruption has no effect since all 'samples' have the same length.

5.3.2 Calculation of the pgf $U_T(z)$

From the two-dimensional pgf $U(z_1, z_2)$, we derive the expression for the marginal pgf of the total system contents:

$$\begin{aligned} U_T(z) \triangleq & \mathrm{E}\left[z^{u_1+u_2}\right] && (5.104)\\ = & U(z,z) && (5.105)\\ = & (1-\rho_{T,eff})\frac{Y_2(z)(z-1)}{z-Y_2(z)} && (5.106)\\ & \times\left\{1+z\frac{(A_T(z)-A(Y_1(z),z))(S_1(A_T(z))-1)}{A(Y_1(z),z)(A_T(z)-1)(z-S_1(A_T(z)))}\right\}. \end{aligned}$$

5.3.3 Calculation of the pgf $U_1(z)$

Furthermore, the pgf of the steady-state class-1 system contents is calculated from $U(z_1, z_2)$:

$$\begin{aligned} U_1(z) \triangleq & \mathrm{E}\left[z^{u_1}\right] && (5.107)\\ = & U(z,1) && (5.108)\\ = & (1-\rho_1)\frac{S_1(A_1(z))(z-1)}{z-S_1(A_1(z))}. && (5.109) \end{aligned}$$

5.3.4 Calculation of the pgf $U_2(z)$

Finally, the pgf of the steady-state class-2 system contents is obtained from $U(z_1, z_2)$ and we find

$$\begin{aligned} U_2(z) \triangleq & \mathrm{E}\left[z^{u_2}\right] && (5.110)\\ = & U(1,z) && (5.111)\\ = & (1-\rho_{T,eff})\frac{A_2(z)(A(Y_1(z),z)-1)}{A(Y_1(z),z)(A_2(z)-1)}\frac{Y_2(z)(z-1)}{z-Y_2(z)}. && (5.112) \end{aligned}$$

5.3.5 Calculation of moments

By taking the necessary derivatives of these (marginal) pgf's, moments of the total, of the class-1 and of the class-2 system contents are found. We show the expressions of the mean values in this subsection.

The mean total system contents is found by taking the first derivative of (5.106) and substituting z by 1:

$$\mathrm{E}[u_T] = U_T'(1) \qquad (5.113)$$

$$
\begin{aligned}
=&\frac{\rho_{T,eff}}{2} + \frac{\mu_1 \text{Var}[a_T]}{2(1-\rho_{T,eff})} - \frac{\mu_1\lambda_2(\mu_{2,eff}-\mu_1)\text{Var}[a_1]}{2(1-\rho_{T,eff})(1-\rho_1)} \\
&+ \frac{(\mu_{2,eff}-\mu_1)\text{Var}[a_2]}{2(1-\rho_{T,eff})} + \frac{\lambda_1\text{Var}[s_1](\lambda_1(1-\rho_1)+\lambda_2(1-\mu_{2,eff}\lambda_1))}{2(1-\rho_{T,eff})(1-\rho_1)} \\
&+ \frac{\lambda_2^2\text{Var}[s_2]_{eff}}{2(1-\rho_{T,eff})(1-\rho_1)} + \frac{\rho_1\lambda_2(\mu_{2,eff}-1)}{2(1-\rho_1)},
\end{aligned} \tag{5.114}
$$

with

$$
\begin{aligned}
\text{Var}[s_2]_{eff} =&\mu_{2,eff}\left[\frac{2(1-\rho_1)A^{(2)}(0,1)}{\lambda_2 A_1(0)(1-A_1(0))}\left\{1-\frac{S_2'(A_1(0))A_1(0)(1-A_1(0))}{S_2(A_1(0))(1-S_2(A_1(0)))}\right\}\right. \\
&\left.+\mu_{2,eff}-1\right],
\end{aligned} \tag{5.115}
$$

for the PRD case, and

$$
\begin{aligned}
\text{Var}[s_2]_{eff} =&\mu_{2,eff}\left[\frac{2(1-\rho_1)A^{(2)}(0,1)}{\lambda_2 A_1(0)(1-A_1(0))}\left\{1-\frac{S_2'(1/A_1(0))}{\mu_{2,eff}}\right\}\right. \\
&\left.+\frac{2(S_2(1/A_1(0)^2)-S_2(1/A_1(0))^2)}{(1/A_1(0)-1)(S_2(1/A_1(0))-1)}+\mu_{2,eff}-1\right],
\end{aligned} \tag{5.116}
$$

for the PRI priority queue.

The mean class-1 system contents is given by

$$
\begin{aligned}
\text{E}[u_1] =&U_1'(1) && (5.117)\\
=&\frac{\rho_1}{2} + \frac{\mu_1\text{Var}[a_1]}{2(1-\rho_1)} + \frac{\lambda_1^2\text{Var}[s_1]}{2(1-\rho_1)}. && (5.118)
\end{aligned}
$$

And finally, the mean class-2 system contents is given by

$$
\begin{aligned}
\text{E}[u_2] =&U_2'(1) && (5.119)\\
=&\frac{\rho_{2,eff}}{2} + \frac{\mu_1^2\lambda_2\text{Var}[a_1]}{2(1-\rho_{T,eff})(1-\rho_1)} + \frac{\mu_{2,eff}\text{Var}[a_2]}{2(1-\rho_{T,eff})} + \frac{\mu_1\text{Cov}[a_1,a_2]}{1-\rho_{T,eff}} && (5.120)\\
&+ \frac{\lambda_2(\lambda_1\text{Var}[s_1]+\lambda_2\text{Var}[s_2]_{eff})}{2(1-\rho_{T,eff})(1-\rho_1)} + \frac{\rho_1\lambda_2(\mu_{2,eff}-1)}{2(1-\rho_1)},
\end{aligned}
$$

with $\text{Var}[s_2]_{eff}$ given by expressions (5.115) or (5.116) for the PRD or PRI priority queue respectively.

It is easily verified that expressions (5.114), (5.118) and (5.120) satisfy the relation $\mathrm{E}[u_T] = \mathrm{E}[u_1] + \mathrm{E}[u_2]$.

5.4 Queue contents

The *queue contents* are easily derived from the system contents. We denote - as before - the queue contents of class-j at the beginning of the k-th slot by $q_{j,k}$ ($j = 1, 2$). The following relation between $q_{j,k}$ and $u_{j,k}$ - the class-j system contents at the beginning of slot k - is then found:

$$q_{1,k} = [u_{1,k} - 1]^+ \tag{5.121}$$

$$q_{2,k} = \begin{cases} [u_{2,k} - 1]^+ & \text{if } u_{1,k} = 0 \\ u_{2,k} & \text{if } u_{1,k} > 0 \end{cases}. \tag{5.122}$$

These are the same expressions as given in the previous chapter (subsection 4.4). When class-1 packets are present in the system, one of them is in service. A class-2 packet is in service when at least one class-2 packet is present and no class-1 packets are present. This leads to the above relations.

Taking the z-transform of these relations, the following relationship between $Q(z_1, z_2)$ - the joint pgf of the steady-state queue contents of class-1 and class-2 at the beginning of a random slot - and $U(z_1, z_2)$ is found

$$Q(z_1, z_2) = \lim_{k \to \infty} \mathrm{E}[z_1^{q_{1,k}} z_2^{q_{2,k}}] \tag{5.123}$$

$$= U(0,0) + \frac{U(0, z_2) - U(0,0)}{z_2} + \frac{U(z_1, z_2) - U(0, z_2)}{z_1}. \tag{5.124}$$

Substituting expression (5.96) in this expression finally yields

$$Q(z_1, z_2) = (1 - \rho_{T,eff}) \frac{z_2 - 1}{z_2 - Y_2(z_2)} \tag{5.125}$$

$$\times \left\{ 1 + \frac{(A(z_1, z_2) - A(Y_1(z_2), z_2))(E_1(z_1, z_2) - 1) Y_2(z_2)}{A(Y_1(z_2), z_2)(A(z_1, z_2) - 1)(z_1 - E_1(z_1, z_2))} \right\}.$$

5.5 Unfinished work

The total unfinished work at the beginning of slot k, denoted by $w_{T,k}$, is defined as the number of slots it takes to serve all packets in the system at the beginning of slot k, when no new packets arrive from slot k on. Furthermore, the unfinished work of class-j ($j = 1, 2$) at the beginning of slot k, denoted by $w_{j,k}$, is defined as the number of slots of the total unfinished work that are effectively spent on serving class-j packets. The steady-state unfinished work

of class-j at the beginning of a random slot is denoted by w_j, $j = 1, 2$. We will perform separate calculations of $W(z_1, z_2)$ - the joint pgf of w_1 and w_2 - for the PRD and PRI priority queues.

PRD

$W(z_1, z_2)$ is expressed as a function of $P(x, z_1, z_2)$ as follows:

$$W(z_1, z_2) \triangleq \mathrm{E}[z_1^{w_1} z_2^{w_2}] \tag{5.126}$$

$$\begin{aligned} =& P(0,0,0) + \frac{P(z_2, 0, S_2(z_2)) - P(0,0,0)}{S_2(z_2)} \\ &+ \frac{P(z_1, S_1(z_1), S_2(z_2)) - P(z_1, 0, S_2(z_2))}{S_1(z_1)}. \end{aligned} \tag{5.127}$$

The first term is the partial pgf when the system is empty at the beginning of the randomly chosen slot. The second term is the partial pgf of the unfinished work of both classes when a class-2 packet is in service at the beginning of the slot. In that case, the remaining service time is part of the class-2 unfinished work. All other class-2 packets in the queue add a class-2 service time to the unfinished work of class-2. The third term is the partial pgf of the unfinished work of both classes when a class-1 packet is in service. In this case, the remaining service time is part of the class-1 unfinished work, and all class-j packets in the queue add a class-j service time to the unfinished work of class-j. Note that the oldest class-2 packet could have already attempted to get served (but was interrupted). Since the priority discipline is of the *PRD* type, this packet indeed also adds a random class-2 service time to the unfinished work.

Substituting expression (5.28) in expression (5.127) gives

$$\begin{aligned} W(z_1, z_2) =& (1 - \rho_{T,eff}) \Bigg\{ 1 + z_1 \frac{(A(S_1(z_1), S_2(z_2)) - A(Y_1(S_2(z_2)), S_2(z_2)))}{A(Y_1(S_2(z_2)), S_2(z_2))(z_1 - A(S_1(z_1), S_2(z_2)))} \\ &\times \frac{Y_2(S_2(z_2))(S_2(z_2) - 1)}{S_2(z_2) - Y_2(S_2(z_2))} + z_2 \frac{A(0, S_2(z_2)) Y_2(S_2(z_2))}{A(Y_1(S_2(z_2)), S_2(z_2))} \\ &\times \frac{(A(Y_1(S_2(z_2)), S_2(z_2)) - 1)(S_2(z_2) - S_2(A(0, S_2(z_2))))}{S_2(A(0, S_2(z_2)))(z_2 - A(0, S_2(z_2)))(S_2(z_2) - Y_2(S_2(z_2)))} \Bigg\}. \end{aligned} \tag{5.128}$$

PRI

In this case, $W(z_1, z_2)$ is written as a function of $P(x, z_1, y_2, z_2)$ as follows:

$$W(z_1, z_2) = \mathrm{E}[z_1^{w_1} z_2^{w_2} \{w_1 = w_2 = 0\}] + \mathrm{E}[z_1^{w_1} z_2^{w_2} \{w_1 = 0, w_2 > 0\}] \tag{5.129}$$

$$
\begin{aligned}
&+ \mathrm{E}[z_1^{w_1} z_2^{w_2} \{w_1 > 0, w_2 = 0\}] + \mathrm{E}[z_1^{w_1} z_2^{w_2} \{w_1 > 0, w_2 > 0\}] \\
=&P(0,0,0,0) + \frac{P(z_2, 0, 1, S_2(z_2)) - P(0,0,0,0)}{S_2(z_2)} \\
&+ \frac{P(z_1, S_1(z_1), 0, 0) - P(0,0,0,0)}{S_1(z_1)} \\
&+ \frac{\left\{ \begin{array}{l} P(z_1, S_1(z_1), z_2, S_2(z_2)) - P(z_1, 0, z_2, S_2(z_2)) \\ -P(z_1, S_1(z_1), 0, 0) + P(0,0,0,0) \end{array} \right\}}{S_1(z_1) S_2(z_2)}.
\end{aligned}
\tag{5.130}
$$

All class-j packets in the queue add a (random) class-j service time to the unfinished work of class-j, except for the (possible) oldest class-1 and class-2 packets. The oldest class-1 packet is being served and it thus adds a residual class-1 service time. The oldest class-2 packet adds a residual service time if no class-1 packets are present at the time. Otherwise it adds a complete service time equal to t_2 (i.e., not just a random class-2 service time)

Substituting expression (5.88) in expression (5.130) gives

$$
\begin{aligned}
W(z_1, z_2) = (1 - \rho_{T,eff}) \Bigg\{ & 1 + z_1 \frac{A(S_1(z_1), S_2(z_2)) - A(Y_1(S_2(z_2)), S_2(z_2))}{A(Y_1(S_2(z_2)), S_2(z_2))(z_1 - A(S_1(z_1), S_2(z_2)))} \\
& \times \frac{B(z_2, S_2(z_2)) - B(1, S_2(z_2))}{S_2(z_2) - B(1, S_2(z_2))} \\
& + \frac{z_2 \left\{ \begin{array}{l} A(0, S_2(z_2)) A(Y_1(S_2(z_2)), S_2(z_2))(S_2(z_2) - B(1, S_2(z_2))) \\ +A(Y_1(S_2(z_2)), S_2(z_2))(B(z_2, S_2(z_2)) - S_2(z_2)) \\ +A(0, S_2(z_2))(B(1, S_2(z_2)) - B(z_2, S_2(z_2))) \end{array} \right\}}{A(Y_1(S_2(z_2)), S_2(z_2))(z_2 - A(0, S_2(z_2)))(S_2(z_2) - B(1, S_2(z_2)))} \Bigg\}.
\end{aligned}
\tag{5.131}
$$

5.6 Packet delay

5.6.1 Pgf $D_1(z)$ of the class-1 packet delay

Since the class-1 characteristics in the preemptive priority queues are independent of whether an interrupted class-2 packet is resumed or repeated, the delays of class-1 packets in preemptive repeat and preemptive resume priority queues are identical. We thus have the same expression for $D_1(z)$ as in the previous chapter (expression (4.136)):

$$
D_1(z) = \frac{1 - \rho_1}{\lambda_1} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1}. \tag{5.132}
$$

5.6.2 Pgf $D_2(z)$ of the class-2 packet delay

We tag a class-2 packet that enters the buffer during slot k. We use the notion of *sub-busy periods* to analyze the class-2 packet delay (as in the previous chapters). Two different kinds of sub-busy periods are defined, i.e., *sub-busy periods initiated by a class-1 packet* and *sub-busy periods initiated by a class-2 packet*. The first type is defined as follows: it starts at the beginning of the slot the initiating class-1 packet enters the server and ends when the number of class-1 packets in the system is one less - for the first time - than when the initiating class-1 packet entered the server. A sub-busy period initiated by a class-2 packet starts at the beginning of the slot the initiating class-2 packet enters the server (for the first time) and it ends at the beginning of which a new class-2 packet can enter the server (if there is any).

Let us refer to the packets in the system at the end of slot k, but that have to be served before the tagged packet as the "primary packets". So, the tagged class-2 packet enters the server for the first time, when *all primary packets* and *all class-1 packets that arrived after slot k* (i.e., while the tagged packet is waiting in the queue) are served. All primary class-j packets (except for the oldest class-1 and class-2 packet) add a class-j sub-busy period to the delay of the tagged packet. Let $\tilde{v}_{j,m}$ denote the length of the sub-busy period added to the tagged packet's delay by the m-th class-j packet already in the queue at the beginning of slot k and let $v_{j,m}^{(i)}$ denote the length of the sub-busy period added to the delay of the tagged class-2 packet by the m-th class-j packet that arrives during slot i. Finally $\tilde{v}_2$ denotes the sub-busy period initiated by the tagged class-2 packet itself.

During the tagged packet's arrival slot, the system is in one of the following states:

1. $u_{1,k} = u_{2,k} = 0$:

$$d_2 = \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k+d_2}} v_{1,m}^{(k+d_2)} \right), \tag{5.133}$$

with $f_{j,k}$ the number of class-j packets arriving during slot k, but that have to be served before the tagged packet. $f_{j,k}$ class-j primary packets ($j = 1, 2$) all add a class-j sub-busy period to d_2. During the service time of the tagged class-2 packet, new class-1 packets may arrive, which interrupt the tagged packet's service. The sub-busy periods added to d_2 by the class-1 packets arriving in the slot preceding the departure of the tagged packet (slot $k + d_2$), are part of $\tilde{v}_2$, but are *not* part of the delay of the tagged packet, since the tagged packet departs from the system at the end of this slot. This accounts for the negative part in the right-hand side of the above expression. Note that the last term in the right-hand side of this expression is stochastically independent of the first term.

2. $u_{1,k} = 0, u_{2,k} > 0$:

$$d_2 = (v_{2,k}^+ - 1) + \sum_{m=1}^{u_{2,k}-1} \tilde{v}_{2,m} + \sum_{m=1}^{f_{2,k}} v_{2,m}^{(k)} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k+d_2}} v_{1,m}^{(k+d_2)} \right), \tag{5.134}$$

with $v_{2,k}^+$ the remaining part of the sub-busy period initiated by the oldest class-2 packet at the beginning of slot k. The difference with the former case, is that (multiple) class-2 packets are present in the system when the tagged class-2 packet arrives and thus have to be served before the tagged packet. All these class-2 packets initiate their own sub-busy periods.

3. $u_{1,k} > 0, u_{2,k} = 0$:

$$\begin{aligned} d_2 = &(r_k - 1) + \sum_{i=1}^{r_k-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{m=1}^{u_{1,k}-1} \tilde{v}_{1,m} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} \\ &+ \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k+d_2}} v_{1,m}^{(k+d_2)} \right). \end{aligned} \tag{5.135}$$

The residual service time of the class-1 packet in service during slot k contributes in the first term, the sub-busy periods added to d_2 by the class-1 packets arriving during the residual service time contribute in the second term, the sub-busy periods added by the class-1 packets already in the queue at the beginning of slot k contribute in the third term, the sub-busy periods added by the class-1 and class-2 packets arriving during slot k, but that have to be served before the tagged class-2 packet contribute in the fourth term and finally the service time of and the sub-busy period initiated by the tagged class-2 packet itself (minus the sub-busy periods initiated by the class-1 packets arriving in the slot preceding the class-2 packet's departure) contribute in the last term.

4. $u_{1,k} > 0, u_{2,k} > 0$:

$$\begin{aligned} d_2 = &(r_k - 1) + \sum_{i=1}^{r_k-1} \sum_{m=1}^{a_{1,k+i}} v_{1,m}^{(k+i)} + \sum_{m=1}^{u_{1,k}-1} \tilde{v}_{1,m} + \sum_{j=1}^{2} \sum_{m=1}^{f_{j,k}} v_{j,m}^{(k)} \\ &+ v_{2,k}^+ + \sum_{m=1}^{u_{2,k}-1} \tilde{v}_{2,m} + \left(\tilde{v}_2 - \sum_{m=1}^{a_{1,k+d_2}} v_{1,m}^{(k+d_2)} \right). \end{aligned} \tag{5.136}$$

This is a combination of the two previous situations.

$D_2(z)$ is given by - by conditioning on whether the class-1 and/or the class-2 systems are empty -

$$\begin{aligned} D_2(z) =& \mathrm{E}\left[z^{d_2}\{u_{1,k}=u_{2,k}=0\}\right] + \mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0\}\right] \\ &+ \mathrm{E}\left[z^{d_2}\{u_{1,k}>0,u_{2,k}=0\}\right] + \mathrm{E}\left[z^{d_2}\{u_{1,k}>0,u_{2,k}>0\}\right]. \end{aligned} \tag{5.137}$$

Before calculating the four terms of expression (5.137) - using the system equations (5.133)-(5.136) - we first take a closer look at the pgf's of the sub-busy periods. It can easily be seen that the sub-busy periods initiated by the primary packets of class-1 (class-2 respectively) form a set of i.i.d. random variables and their common pgf is - as usual - represented by $V_1(z)$ ($V_2(z)$ respectively).

The sub-busy periods initiated by a class-1 packet are still identically distributed as in the previous chapters. Thus we have

$$V_1(z) = S_1(zA_1(V_1(z))). \tag{5.138}$$

The sub-busy periods initiated by a class-2 packet are different from the ones discussed in the previous chapters. The reason for this is that the possible repeats of the class-2 packet's service time have to be accounted for in the sub-busy period in this case. These sub-busy periods are furthermore different for the PRD and PRI priority queues.

From this point forward we continue the analyses of the PRD and PRI priority queues separately. We start with the PRD case.

PRD

We first calculate $V_2(z)$. When a class-2 packet enters the server, two events can occur: the class-2 packet can either be completely served in the first attempt, or the service of the class-2 packet is interrupted by newly arriving class-1 packets during its service time. Denoting a random class-2 sub-busy period by v_2, we find

$$V_2(z) \triangleq \mathrm{E}[z^{v_2}] \tag{5.139}$$

$$= \mathrm{E}[z^{v_2}\{\text{no interruption}\}] + \mathrm{E}[z^{v_2}\{\text{interruption}\}], \tag{5.140}$$

with "interruption" and "no interruption" short for the events that the first attempt of the class-2 service is interrupted or not interrupted respectively by arriving class-1 packets. A class-2 service is not interrupted if no class-1 packets arrive during the slots of the class-2 service time, with exception of the last slot (since the class-2 packet leaves the system at the end of that slot independent of whether class-1 packets arrive during that slot). The possible class-1 packets that enter the system during that last service slot add class-1

sub-busy periods to the class-2 sub-busy period. Therefore the first term of (5.140) is given by

$$\mathrm{E}[z^{v_2}\{\text{no interruption}\}] = A_1(V_1(z)) \sum_{i=1}^{\infty} s_2(i) A_1(0)^{i-1} z^i \tag{5.141}$$

$$= \frac{S_2(zA_1(0))A_1(V_1(z))}{A_1(0)}, \tag{5.142}$$

with $s_2(i)$ the pmf of the class-2 service times.

The class-2 service is interrupted when $m \geq 1$ class-1 packets arrive during one of the service slots of the class-2 packet (excluding the last one). These class-1 packets all add a class-1 sub-busy period. The class-2 packet is put back in the queue and has to wait in the queue until all class-1 packets are served before having another service attempt. Since the service time in the new attempt is resampled, this again initiates a new class-2 sub-busy period with pgf $V_2(z)$. This leads to the following expression

$$\mathrm{E}[z^{v_2}\{\text{interruption}\}] = V_2(z) \sum_{i=1}^{\infty} s_2(i) \sum_{j=1}^{i-1} A_1(0)^{j-1} z^j \sum_{m=1}^{\infty} a_1(m)(V_1(z))^m \tag{5.143}$$

$$= \frac{(A_1(V_1(z)) - A_1(0))(S_2(A_1(0)z) - A_1(0)z)V_2(z)}{A_1(0)(A_1(0)z - 1)}, \tag{5.144}$$

with $a_1(m)$ the pmf of the per-slot class-1 arrivals.

Substituting expressions (5.142) and (5.144) in (5.140), we find the following expression for $V_2(z)$:

$$V_2(z) = \frac{(1 - A_1(0)z)A_1(V_1(z))S_2(A_1(0)z)}{(A_1(V_1(z)) - A_1(0))S_2(A_1(0)z) - A_1(0)(zA_1(V_1(z)) - 1)}. \tag{5.145}$$

We now return to formula (5.137) and calculate the four partial pgf's one by one. Firstly, z-transforming expression (5.133), we find

$$\mathrm{E}\left[z^{d_2}\{u_{1,k} = u_{2,k} = 0\}\right] = \frac{F^{(2)}(V_1(z), V_2(z))V_2(z)P(0,0,0)}{A_1(V_1(z))}, \tag{5.146}$$

with

$$F^{(2)}(z_1, z_2) \triangleq \mathrm{E}\left[z_1^{f_{1,k}} z_2^{f_{2,k}}\right] \tag{5.147}$$

$$= \frac{A(z_1, z_2) - A_1(z_1)}{\lambda_2(z_2 - 1)}, \tag{5.148}$$

which was already calculated in the previous chapters.

The second term in expression (5.137) is a bit more involved, because in this case - $u_{1,k} = 0, u_{2,k} > 0$ - a class-2 packet is in service when the tagged packet arrives, which can be interrupted by newly arriving class-1 packets. Thus, the packet in service during slot k is either completely served during that service attempt or is interrupted by class-1 packets, leading to:

$$\begin{aligned}\mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0\}\right] =& \mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0,\text{no interruption}\}\right] \\ &+ \mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0,\text{interruption}\}\right].\end{aligned} \tag{5.149}$$

The first term equals - by z-transforming expression (5.134) -

$$\begin{aligned}&\mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0,\text{no interruption}\}\right] \\ &\quad=\frac{F^{(2)}(V_1(z),V_2(z))V_2(z)}{A_1(V_1(z))}R_2(V_2(z)) + \frac{F^{(2)}(0,V_2(z))}{(A_1(0))^2 z} \\ &\qquad\times[P(A_1(0)z,0,V_2(z)) - A_1(0)zV_2(z)R_2(V_2(z)) - P(0,0,0)],\end{aligned} \tag{5.150}$$

with $R_2(z_2)$ and $P(x,z_1,z_2)$ defined in section 5.3 The first term is the partial pgf of when the remaining service time of the class-2 packet equals 1 slot, while the second term gives the partial pgf when this remaining service time is larger than 1. Note that in this latter case $f_{1,k} = 0$, since we consider the case that the service of the packet in service during slot k is not interrupted.

The second term of (5.149) is given by - again by using expression (5.134) and after some extensive mathematical manipulations -

$$\begin{aligned}&\mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0,\text{interruption}\}\right] \\ &=(F^{(2)}(V_1(z),V_2(z)) - F^{(2)}(0,V_2(z)))[P(1,0,V_2(z)) - V_2(z)R_2(V_2(z)) \\ &\quad - P(0,0,0)]\frac{V_2(z)}{A_1(V_1(z))} + F^{(2)}(0,V_2(z))\frac{z(A_1(V_1(z)) - A_1(0))}{A_1(0)z-1} \\ &\quad\times\left\{\frac{P(A_1(0)z,0,V_2(z)) - A_1(0)zV_2(z)R_2(V_2(z)) - P(0,0,0)}{(A_1(0)z)^2}\right. \\ &\quad\left. - [P(1,0,V_2(z)) - V_2(z)R_2(V_2(z)) - P(0,0,0)]\right\}\frac{V_2(z)}{A_1(V_1(z))}.\end{aligned} \tag{5.151}$$

The first term of this expression is the partial pgf when class-1 packets arrive during slot k, while the second term is the partial pgf when no class-1 arrivals occur during this slot.

Thirdly, we take a look at the third and fourth term of expression (5.137). Note, that $v_{2,k}^+$ is equally distributed as a *complete* sub-busy period when $u_{1,k} > 0, u_{2,k} > 0$ since the oldest class-2 packet is waiting in the queue at the beginning of slot k and as a result it starts a "new" complete sub-busy period

with pgf $V_2(z)$ once it (re)starts service. (This is due to the resampling of the service time after an interruption and will thus not be valid for the PRI priority queue.) As a result, it can be seen from expressions (5.135)-(5.136) that the situations $u_{1,k} > 0, u_{2,k} = 0$ and $u_{1,k} > 0, u_{2,k} > 0$ can be analyzed together. We find

$$\mathrm{E}\left[z^{d_2}\{u_{1,k} > 0\}\right] = \mathrm{E}\left[z^{d_2}\{u_{1,k} > 0, u_{2,k} = 0\}\right] + \mathrm{E}\left[z^{d_2}\{u_{1,k} > 0, u_{2,k} > 0\}\right] \tag{5.152}$$

$$= \frac{V_2(z)F^{(2)}(V_1(z), V_2(z))}{A_1(V_1(z))} \tag{5.153}$$
$$\times \frac{P(zA_1(V_1(z)), V_1(z), V_2(z)) - P(zA_1(V_1(z)), 0, V_2(z))}{zA_1(V_1(z))V_1(z)}.$$

Summing expressions (5.146), (5.150), (5.151) and (5.153) yields the following expression for $D_2(z)$

$$\begin{aligned} D_2(z) = &\frac{V_2(z)}{A_1(V_1(z))}\Bigg\{F^{(2)}(V_1(z), V_2(z))[P(0,0,0) + R_2(V_2(z))] \\ &+ \left[F^{(2)}(V_1(z), V_2(z)) - F^{(2)}(0, V_2(z))\frac{zA_1(V_1(z)) - 1}{A_1(0)z - 1}\right][P(1,0,V_2(z)) \\ &- V_2(z)R_2(V_2(z)) - P(0,0,0)] + F^{(2)}(0, V_2(z))\frac{zA_1(V_1(z)) - 1}{A_1(0)z - 1} \\ &\times \frac{P(A_1(0)z, 0, V_2(z)) - A_1(0)zV_2(z)R_2(V_2(z)) - P(0,0,0)}{A_1(0)zS_2(A_1(0)z)} \\ &+ F^{(2)}(V_1(z), V_2(z)) \\ &\times \frac{P(zA_1(V_1(z)), V_1(z), V_2(z)) - P(zA_1(V_1(z)), 0, V_2(z))}{zA_1(V_1(z))V_1(z)}\Bigg\}. \end{aligned} \tag{5.154}$$

Using expressions (5.26), (5.28), (5.29) and (5.148) in the previous expression of $D_2(z)$, we finally find

$$\begin{aligned} D_2(z) = &\frac{1 - \rho_{T,eff}}{\lambda_2}\frac{V_2(z)}{A_1(V_1(z))}\Bigg\{\frac{(zA_1(V_1(z)) - 1)(A(0, V_2(z)) - A_1(0))}{(V_2(z) - 1)(A_1(0)z - A(0, V_2(z)))} \\ &+ \frac{Y_2(V_2(z))(zA_1(V_1(z)) - A(Y_1(V_2(z)), V_2(z)))}{A(Y_1(V_2(z)), V_2(z))(V_2(z) - Y_2(V_2(z)))} \\ &\times \frac{(A_1(0)A(V_1(z), V_2(z)) - A_1(V_1(z))A(0, V_2(z)))(z-1)}{(A_1(0)z - A(0, V_2(z)))(zA_1(V_1(z)) - A(V_1(z), V_2(z)))}\Bigg\}, \end{aligned} \tag{5.155}$$

with $Y_2(z)$ given by (5.27).

PRI

We first calculate $V_2(z)$ for this case. We denote the conditional pgf of the sub-busy period of a class-2 packet with a service time of i slots by $V_{2,i}(z)$. We thus have

$$V_2(z) = \sum_{i=1}^{\infty} s_2(i) V_{2,i}(z). \tag{5.156}$$

We first calculate an expression for the $V_{2,i}(z)$. The pgf of the sub-busy period initiated by a class-2 service time equal to i slots is the same as the pgf of the class-2 sub-busy period in the case of a *PRD* priority queue and *deterministic* class-2 service times of i slots. Indeed, in case of deterministic service times resampling or not resampling are identical. This $V_{2,i}(z)$ thus equals expression (5.145) with $S_2(z) = z^i$, or

$$V_{2,i}(z) = \frac{(1 - A_1(0)z)A_1(V_1(z))(A_1(0)z)^i}{(A_1(V_1(z)) - A_1(0))(A_1(0)z)^i - A_1(0)(zA_1(V_1(z)) - 1)}. \tag{5.157}$$

Substituting this expression in (5.156) yields

$$V_2(z) = \sum_{i=1}^{\infty} \frac{s_2(i)(1 - A_1(0)z)A_1(V_1(z))(A_1(0)z)^i}{(A_1(V_1(z)) - A_1(0))(A_1(0)z)^i - A_1(0)(zA_1(V_1(z)) - 1)}. \tag{5.158}$$

We calculate the four partial pgf's in the right-hand side of expression (5.137) one by one in the PRI case. Firstly, taking the z-transform of equation (5.133) yields

$$\mathrm{E}\left[z^{d_2}\{u_{1,k} = u_{2,k} = 0\}\right] = \frac{F^{(2)}(V_1(z), V_2(z))V_2(z)P(0,0,0,0)}{A_1(V_1(z))}, \tag{5.159}$$

with $F^{(2)}(z_1, z_2)$ still given by expression (5.148).

The second term in expression (5.137) is (again) a bit more involved. The class-2 packet in service during slot k is either served completely during that transmission attempt or is interrupted by class-1 packets, leading to:

$$\begin{aligned}\mathrm{E}\left[z^{d_2}\{u_{1,k} = 0, u_{2,k} > 0\}\right] =& \mathrm{E}\left[z^{d_2}\{u_{1,k} = 0, u_{2,k} > 0, \text{no interruption}\}\right] \\ &+ \mathrm{E}\left[z^{d_2}\{u_{1,k} = 0, u_{2,k} > 0, \text{interruption}\}\right].\end{aligned} \tag{5.160}$$

The first term equals - by z-transforming expression (5.134) -

$$\mathrm{E}\left[z^{d_2}\{u_{1,k} = 0, u_{2,k} > 0, \text{no interruption}\}\right] \tag{5.161}$$

$$=\frac{F^{(2)}(V_1(z),V_2(z))V_2(z)}{A_1(V_1(z))}R_2(1,V_2(z))+\frac{F^{(2)}(0,V_2(z))}{(A_1(0))^2z}$$
$$\times[P(A_1(0)z,0,1,V_2(z))-A_1(0)zV_2(z)R_2(1,V_2(z))-P(0,0,0,0)].$$

The first term is the partial pgf when the remaining service time of the class-2 packet in service during slot k equals 1 slot, while the second term yields the partial pgf when this remaining service time is larger than 1.

The second term of (5.160) is given by - by using expression (5.134) and after some mathematical manipulations

$$\mathrm{E}\left[z^{d_2}\{u_{1,k}=0,u_{2,k}>0,\text{interruption}\}\right] \tag{5.162}$$
$$=\frac{F^{(2)}(V_1(z),V_2(z))-F^{(2)}(0,V_2(z))}{A_1(V_1(z))}\sum_{i=1}^{\infty}V_{2,i}(z)\left[Q_i(1,0,V_2(z))\right.$$
$$\left.-s_2(i)V_2(z)R_{2,i}(V_2(z))\right]+\frac{F^{(2)}(0,V_2(z))}{A_1(V_1(z))}\frac{z(A_1(V_1(z)))-A_1(0)}{A_1(0)z-1}$$
$$\times\left\{\frac{\sum_{i=1}^{\infty}V_{2,i}(z)[Q_i(A_1(0)z,0,V_2(z))-s_2(i)A_1(0)zV_2(z)R_{2,i}(V_2(z))]}{(A_1(0)z)^2}\right.$$
$$\left.-\sum_{i=1}^{\infty}V_{2,i}(z)\left[Q_i(1,0,V_2(z))-s_2(i)V_2(z)R_{2,i}(V_2(z))\right]\right\},$$

with

$$Q_i(x,z_1,z_2)\triangleq\mathrm{E}[x^{r_k}z_1^{u_{1,k}}z_2^{u_{2,k}}\{t_{2,k}=i\}], \tag{5.163}$$

$i>0$. We show later on how the latter partial pgf's are calculated using the expression of $P(x,z_1,y_2,z_2)$. The first term of expression (5.162) is the partial pgf when class-1 packets arrive during slot k, while the second term is the partial pgf when no class-1 arrivals occur during this slot. Note that expression (5.162) resembles expression (5.151) in the PRD case. The difference is that in the PRI case, we have to condition on the service time t_2 of the class-2 packet which service is preempted because this service time is *not* resampled in this case.

Furthermore, we take a look at the third term of expression (5.137). We find

$$\mathrm{E}\left[z^{d_2}\{u_{1,k}>0,u_{2,k}=0\}\right]=\frac{F^{(2)}(V_1(z),V_2(z))V_2(z)}{A_1(V_1(z))} \tag{5.164}$$
$$\times\frac{P(zA_1(V_1(z)),V_1(z),0,0)-P(0,0,0,0)}{zA_1(V_1(z))V_1(z)}.$$

Finally, the last term of expression (5.137) is given by

$$\mathrm{E}\big[z^{d_2}\{u_{1,k}>0,u_{2,k}>0\}\big] = \frac{F^{(2)}(V_1(z),V_2(z))}{A_1(V_1(z))} \tag{5.165}$$
$$\times \frac{\sum_{i=1}^{\infty} V_{2,i}(z)[Q_i(zA_1(V_1(z)),V_1(z),V_2(z)) - Q_i(zA_1(V_1(z)),0,V_2(z))]}{zA_1(V_1(z))V_1(z)}.$$

Again, we have to condition on the service time of the oldest class-2 packet because this service time could have been interrupted before.

Summing expressions (5.159), (5.161), (5.162), (5.164) and (5.165) yields an expression for $D_2(z)$ in terms of the functions $P(.,.,.,.)$ and $R_2(1,.)$ and the functions $R_{2,i}(.)$ and $Q_i(.,.,.)$ $(i>0)$. The $R_{2,i}(z_2)$, $R_2(1,z_2)$ and $P(x,z_1,y_2,z_2)$ are given by expressions (5.83), (5.86) and (5.88) respectively. So it remains for us to calculate the $Q_i(x,z_1,z_2)$. From the definition (5.163) of $Q_i(x,z_1,z_2)$ and $P(x,z_1,y_2,z_2)$, it is seen that the following relation exists:

$$P(x,z_1,y_2,z_2) - P(x,z_1,0,0) = \sum_{i=1}^{\infty} Q_i(z_1,y_2,z_2)y_2^i. \tag{5.166}$$

From expressions (5.75) and (5.88), we find

$$P(x,z_1,y_2,z_2) - P(x,z_1,0,0) = (1-\rho_{T,eff})x \tag{5.167}$$
$$\times \left[z_1 \frac{(A(z_1,0)-A(Y_1(0),0))(E_1(z_1,0)-S_1(x))S_2(y_2)}{A(Y_1(0),0)(x-A(z_1,0))(z_1-E_1(z_1,0))} \right.$$
$$+ z_1 \frac{(A(z_1,z_2)-A(Y_1(z_2),z_2))(S_1(x)-E_1(z_1,z_2))}{A(Y_1(z_2),z_2)(x-A(z_1,z_2))(z_1-E_1(z_1,z_2))(z_2-B(1,z_2))}$$
$$\times (z_2B(y_2,z_2) - S_2(y_2)B(1,z_2))$$
$$\left. + z_2 \frac{\left\{ \begin{array}{l} A(Y_1(z_2),z_2)(B(xy_2,z_2)-S_2(xy_2)) \\ +A(0,z_2)A(Y_1(z_2),z_2)(S_2(xy_2)-B(y_2,z_2)) \\ +A(0,z_2)(B(y_2,z_2)-B(xy_2,z_2)) \end{array} \right\}}{A(Y_1(z_2),z_2)(x-A(0,z_2))(z_2-B(1,z_2))} \right].$$

The right-hand sides of the last two expressions have to be equal for each y_2, $|y_2|<1$. Thus, the coefficients of the y_2^i of both expressions are also equal (for all i), leading to - by noting that $B(y,z)$ is given by (5.85) and $S_2(y) = \sum_{i=1}^{\infty} s_2(i)y^i$ -

$$Q_i(x,z_1,z_2) = \tag{5.168}$$
$$(1-\rho_{T,eff})xs_2(i) \left[\frac{z_1(A(z_1,0)-A(Y_1(0),0))(E_1(z_1,0)-S_1(x))}{A(Y_1(0),0)(x-A(z_1,0))(z_1-E_1(z_1,0))} \right.$$

$$+\frac{z_1(A(z_1,z_2)-A(Y_1(z_2),z_2))(S_1(x)-E_1(z_1,z_2))}{A(Y_1(z_2),z_2)(x-A(z_1,z_2))(z_1-E_1(z_1,z_2))(z_2-B(1,z_2))}$$
$$\times\left\{\frac{z_2(1-A(0,z_2))A(Y_1(z_2),z_2)A(0,z_2)^i}{(A(Y_1(z_2),z_2)-A(0,z_2))A(0,z_2)^i-A(0,z_2)(A(Y_1(z_2),z_2)-1)}\right.$$
$$\left.-B(1,z_2)\right\}+\frac{z_2(x^i-A(0,z_2)^i)}{(A(Y_1(z_2),z_2)-A(0,z_2))A(0,z_2)^{i-1}-(A(Y_1(z_2),z_2)-1)}$$
$$\times\frac{(1-A(0,z_2))(A(Y_1(z_2),z_2)-1)}{(x-A(0,z_2))(z_2-B(1,z_2))}\Bigg].$$

Finally bringing everything together, we find (after some extensive mathematical manipulations)

$$D_2(z)=\frac{1-\rho_{T,eff}}{\lambda_2}\frac{V_2(z)}{A_1(V_1(z))}\left\{\frac{(zA_1(V_1(z))-1)(A(0,V_2(z))-A_1(0))}{(V_2(z)-1)(A_1(0)z-A(0,V_2(z)))}\right.$$
$$+\frac{(zA_1(V_1(z))-A(Y_1(V_2(z)),V_2(z)))\sum_{i=1}^{\infty}s_2(i)(V_{2,i}(z)-1)Y_{2,i}(V_2(z))}{A(Y_1(V_2(z)),V_2(z))(V_2(z)-Y_2(V_2(z)))(V_2(z)-1)}$$
$$\left.\times\frac{(A_1(0)A(V_1(z),V_2(z))-A_1(V_1(z))A(0,V_2(z)))(z-1)}{(A_1(0)z-A(0,V_2(z)))(zA_1(V_1(z))-A(V_1(z),V_2(z)))}\right\}, \tag{5.169}$$

with $Y_2(z)$ given by (5.99) and

$$Y_{2,i}(z)\triangleq\frac{(1-A(0,z))A(Y_1(z),z)A(0,z)^i}{(A(Y_1(z),z)-A(0,z))A(0,z)^i-A(0,z)(A(Y_1(z),z)-1)}, \tag{5.170}$$

or thus

$$Y_2(z)=\sum_{i=1}^{\infty}s_2(i)Y_{2,i}(z). \tag{5.171}$$

Special case: geometric class-2 service times

In the special case of geometric service times of class-2, i.e.,

$$S_2(z)=\frac{(1-\beta_2)z}{1-\beta_2 z}, \tag{5.172}$$

expression (5.155) equals

$$D_2(z)=\frac{1-\rho_T}{\rho_2}\frac{z(A(V_1(z),V_2(z))-A_1(V_1(z)))}{zA_1(V_1(z))-A(V_1(z),V_2(z))}, \tag{5.173}$$

with $V_2(z) = S_2(zA_1(V_1(z)))$ in the case of PRD. This is the same expression as found in chapter 4 for a preemptive *resume* priority queue with the same arrival and service processes. As discussed earlier, a remaining class-2 service time and (a new sample of) a complete class-2 service time have the same geometric distribution in this case, and thus both PR and PRD queues are stochastically equal.

Special case: deterministic class-2 service times

In the special case of deterministic class-2 service times, i.e., $S_2(z) = z^{\mu_2}$ and $s_2(i) = \mu_2$, we find in the PRI case that

$$\frac{\sum_{i=1}^{\infty} s_2(i)(V_{2,i}(z) - 1)Y_{2,i}(V_2(z))}{(V_2(z) - Y_2(V_2(z)))(V_2(z) - 1)} = \frac{Y_2(V_2(z))}{V_2(z) - Y_2(V_2(z))}. \tag{5.174}$$

Since, $Y_2(z)$ and $V_2(z)$ are - in this special case - equal for the PRD and PRI case, $D_2(z)$ also becomes equal for both priority queues. Indeed, as already mentioned, resampling or not resampling are identical when the class-2 service times are of constant length.

Special case: uncorrelated number of per-slot class-1 and class-2 arrivals

In this case $A(z_1, z_2) = A_1(z_1)A_2(z_2)$. The second terms of the right-hand sides of expressions (5.155) and (5.169) are equal to zero in this case and $D_2(z)$ equals

$$D_2(z) = \frac{1 - \rho_{T,eff}}{\lambda_2} \frac{V_2(z)(zA_1(V_1(z)) - 1)}{A_1(V_1(z))(V_2(z) - 1)} \frac{A_2(V_2(z)) - 1}{z - A_2(V_2(z))}. \tag{5.175}$$

Note that in the case of PR, the same expression is obtained in case of uncorrelated per-slot class-1 and class-2 arrivals (see expression (4.143)). So, in this case, the distribution of class-2 sub-busy periods is still different for the three priority queues (PR, PRD and PRI), but the relationship between the delay of a class-2 packet and the sub-busy periods of class-2 packets is identical for the three cases. In Fiems [2004], a similar relationship is found. In Fiems' dissertation, the low-priority characteristics of priority queues are analyzed by using queues with server interruptions. Indeed, from the point-of-view of class-2 packets the server is interrupted when class-1 packets are being served. In Fiems [2004], these interruptions are incorporated in the service times - and are called *effective* service times - which are analyzed separately for the PR, PRD and PRI queue respectively. Once the distributions of these effective service times are found however, the PR, PRD and PRI queues are analyzed in a uniform manner. The effective service times in that dissertation are thus

closely related to the sub-busy periods initiated by class-2 packets defined in our dissertation.

Note that - as expressions (5.155) and (5.169) show - $D_2(z)$ is much more complicated in the case the numbers of per-slot class-1 and class-2 arrivals are correlated.

5.6.3 Pgf $D(z)$ of the delay of a random packet

We tag a random (class-1 or class-2) packet. Since, the probability that the tagged packet is of class-j is equal to λ_j/λ_T, we find for the pgf of the delay of a random packet

$$D(z) = \frac{\lambda_1}{\lambda_T} D_1(z) + \frac{\lambda_2}{\lambda_T} D_2(z) \tag{5.176}$$

$$= \frac{1-\rho_1}{\lambda_T} \frac{S_1(z)(z-1)}{z - A_1(S_1(z))} \frac{A_1(S_1(z)) - 1}{S_1(z) - 1} \tag{5.177}$$

$$\frac{1-\rho_{T,eff}}{\lambda_T} \frac{V_2(z)}{A_1(V_1(z))} \left\{ \frac{(zA_1(V_1(z)) - 1)(A(0, V_2(z)) - A_1(0))}{(V_2(z) - 1)(A_1(0)z - A(0, V_2(z)))} \right.$$

$$+ \frac{(zA_1(V_1(z)) - A(Y_1(V_2(z)), V_2(z)))C(z)}{A(Y_1(V_2(z)), V_2(z))}$$

$$\left. \times \frac{(A_1(0)A(V_1(z), V_2(z)) - A_1(V_1(z))A(0, V_2(z)))(z-1)}{(A_1(0)z - A(0, V_2(z)))(zA_1(V_1(z)) - A(V_1(z), V_2(z)))} \right\},$$

with

$$C(z) \triangleq \frac{Y_2(V_2(z))}{V_2(z) - Y_2(V_2(z))}, \tag{5.178}$$

in case of PRD, and

$$C(z) = \frac{\sum_{i=1}^{\infty} s_2(i)(V_{2,i}(z) - 1)Y_{2,i}(V_2(z))}{(V_2(z) - Y_2(V_2(z)))(V_2(z) - 1)}, \tag{5.179}$$

in case of PRI.

5.6.4 The functions $Y_j(z)$

As in the previous chapters $Y_j(z)$ ($j = 1, 2$) is the pgf of the number of class-2 arrivals during a sub-busy period initiated by a class-j packet. For $Y_1(z)$ this is directly clear, since $Y_1(z)$ is identically defined as in the previous chapters. In case of $Y_2(z)$ on the other hand, the possible repeats of the class-2 service times have to be taken into account. Therefore we will in this section calculate

the pgf of the number of class-2 arrivals during a sub-busy period initiated by a class-2 packet and prove that it equals $Y_2(z)$.

When at the beginning of a slot a class-2 packet with service time s_2 enters the server, a new sub-busy period starts (with length denoted by v_2). Denoting the number of class-2 packets that arrive during this sub-busy period by y_2 and denoting the number of class-j arrivals during the i-th slot of this sub-busy period by $a_j^{(i)}$ ($j = 1, 2$), we get

$$y_2 = \sum_{i=1}^{v_2} a_2^{(i)}. \tag{5.180}$$

We prove in the remainder of this subsection, that $Y_2(z)$ is the pgf of y_2 for both the PRD and the PRI queue.

PRD

When a class-2 packet enters the server, two events may occur: the class-2 packet is either completely served in the first attempt, or the service of the class-2 packet is interrupted by newly arriving class-1 packets. The service time of the class-2 packet is not interrupted if no class-1 packets arrive during all slots of the class-2 service time (with exception of its last slot). The possible class-1 packets that enter the system during that last service slot add class-1 sub-busy periods to the the (initial) class-2 sub-busy period and thus the class-2 arrivals during those class-1 sub-busy period are part of y_2. Since $Y_1(z)$ is the pgf of the number of class-2 arrivals during a class-1 sub-busy period, the partial pgf of y_2 in case of no interruptions during the initial attempt of the class-2 packet to get transmitted is given by

$$\mathrm{E}[z^{y_2}\{\text{no interruption}\}] = \frac{S_2(A(0,z))}{A(0,z)} A(Y_1(z), z). \tag{5.181}$$

The first factor is the pgf of the number of class-2 arrivals during all slots of the initial service time (excluding the last slot) and the second factor is the pgf of the number of class-2 arrivals during the last service slot and during the sub-busy periods added by the class-1 arrivals in this last slot.

The class-2 service is interrupted before completing its service time when $m \geq 1$ class-1 packets arrive during one of the service slots of the class-2 packet (excluding the last one). These class-1 packets all initiate a class-1 sub-busy period. The class-2 packet is put back in the queue and has to wait in the queue until all class-1 packets are served before having another service attempt. Since the service time in the new attempt is resampled, this again initiates a new class-2 sub-busy period and thus the class-2 arrivals during this

new sub-busy period - denoted by y_2^* - is equally distributed as y_2. This leads to the following expression (in a similar way as reaching expression (5.144)):

$$\mathrm{E}[z^{y_2}\{\text{interruption}\}] = \frac{(A(Y_1(z),z) - A(0,z))(S_2(A(0,z)) - A(0,z))\mathrm{E}[z^{y_2^*}]}{A(0,z)(A(0,z)-1)}. \tag{5.182}$$

From expressions (5.181) and (5.182) and from the fact that y_2 and y_2^* are equally distributed we indeed find expression (5.27) for the pgf of y_2.

PRI

In this case, we condition on the lengths of the service times. Thus

$$\mathrm{E}[z^{y_2}] = \sum_{i=1}^{\infty} s_2(i)\mathrm{E}[z^{y_2}|s_2 = i], \tag{5.183}$$

with s_2 the service time of the initial class-2 packet. $\mathrm{E}[z^{y_2}|s_2 = i]$ equals $Y_2(z)$ in case of PRD and of deterministic class-2 service times of i slots, i.e.,

$$\mathrm{E}[z^{y_2}|s_2 = i] = \frac{(1 - A(0,z))A(Y_1(z),z)A(0,z)^i}{(A(Y_1(z),z) - A(0,z))A(0,z)^i - A(0,z)(A(Y_1(z),z) - 1)}. \tag{5.184}$$

Substituting this in the previous expression, we indeed find expression (5.99).

When the number of class-1 and class-2 arrivals are independent stochastic variables, it can be seen that v and the $a_2^{(i)}$ are also independent variables. From expression (5.180), it then follows that

$$Y_2(z) = V_2(A_2(z)). \tag{5.185}$$

Indeed, by substituting $A(z_1, z_2)$ by $A_1(z_1)A_2(z_2)$, expression (5.27) (expression (5.99)) and expression (5.145) (expression (5.158)) satisfy the previous relationship in case of PRD (PRI respectively), since $Y_1(z) = V_1(A_2(z))$ in this case.

Note that expression (5.185) is not generally valid when the number of class-1 and class-2 arrivals in a slot are correlated, since v_2 and the $a_2^{(i)}$ both depend on the $a_1^{(i)}$ (in general).

5.6.5 Calculation of moments

The mean class-1 packet delay is given by

$$\mathrm{E}[d_1] = D_1'(1) \tag{5.186}$$

$$= \frac{\mu_1}{2} + \frac{\mu_1 \mathrm{Var}[a_1]}{2\lambda_1(1-\rho_1)} + \frac{\lambda_1 \mathrm{Var}[s_1]}{2(1-\rho_1)}. \tag{5.187}$$

The mean class-2 packet delay is expressed as follows:

$$\mathrm{E}[d_2] = D_2'(1) \tag{5.188}$$

$$= \frac{\mu_{2,eff}}{2} + \frac{\mu_1^2 \mathrm{Var}[a_1]}{2(1-\rho_{T,eff})(1-\rho_1)} + \frac{\mu_{2,eff}\mathrm{Var}[a_2]}{2\lambda_2(1-\rho_{T,eff})} + \frac{\mu_1 \mathrm{Cov}[a_1,a_2]}{\lambda_2(1-\rho_{T,eff})} \tag{5.189}$$

$$+ \frac{(\lambda_1 \mathrm{Var}[s_1] + \lambda_2 \mathrm{Var}[s_2]_{eff})}{2(1-\rho_{T,eff})(1-\rho_1)} + \frac{\rho_1(\mu_{2,eff}-1)}{2(1-\rho_1)},$$

with $\mathrm{Var}[s_2]_{eff}$ given by expressions (5.115) or (5.116) for the PRD or PRI priority queue respectively.

Finally, the mean delay of a random packet yields

$$\mathrm{E}[d] = D'(1) \tag{5.190}$$

$$= \frac{\rho_{T,eff}}{2\lambda_T} + \frac{\mu_1 \mathrm{Var}[a_T]}{2\lambda_T(1-\rho_{T,eff})} - \frac{\mu_1\lambda_2(\mu_{2,eff}-\mu_1)\mathrm{Var}[a_1]}{2\lambda_T(1-\rho_{T,eff})(1-\rho_1)} \tag{5.191}$$

$$+ \frac{(\mu_{2,eff}-\mu_1)\mathrm{Var}[a_2]}{2\lambda_T(1-\rho_{T,eff})} + \frac{\lambda_1 \mathrm{Var}[s_1](\lambda_1(1-\rho_1)+\lambda_2(1-\mu_{2,eff}\lambda_1))}{2\lambda_T(1-\rho_{T,eff})(1-\rho_1)}$$

$$+ \frac{\lambda_2^2 \mathrm{Var}[s_2]_{eff}}{2\lambda_T(1-\rho_{T,eff})(1-\rho_1)} + \frac{\rho_1\lambda_2(\mu_{2,eff}-1)}{2\lambda_T(1-\rho_1)}.$$

It can be seen from the expressions in this subsection and subsection 5.3.5 that Little's law holds for the total system and for the class-1 and class-2 systems separately.

5.7 Waiting time

The waiting time is defined as the number of slots a packet has to wait in the *queue* before starting service. Thus specifically for the class-2 packets, the waiting time - as defined in this dissertation - does not include the slots the packets spend in the queue after the possible interruption(s).

We find the same expression for the pgf $T_1(z)$ as in the previous chapter since the class-1 waiting time is independent of whether service times of class-2 packets are resumed or repeated after an interruption. Therefore, we have

$$T_1(z) = \frac{1-\rho_1}{\lambda_1} \frac{z-1}{z - A_1(S_1(z))} \frac{A_1(S_1(z))-1}{S_1(z)-1} \tag{5.192}$$

for the pgf of the class-1 waiting time.

The relation between the waiting time t_2 and the delay d_2 of a class-2 packet is given by

$$d_2 = t_2 + \tilde{v}_2 - \sum_{m=1}^{a_{1,k+d_2}} v_{1,m}^{(k+d_2)}, \tag{5.193}$$

with $\tilde{v}_2$ the sub-busy period initiated by the class-2 packet and $v_{1,m}^{(k+d_2)}$ the sub-busy periods added to the delay by the class-1 packets arriving during the last service slot of the class-2 packet. The class-1 packets arriving during the tagged class-2 packet's last service slot are part of the sub-busy period initiated by the class-2 packet but are not part of its delay (since the class-2 packet leaves the system at the end of that slot). Since the first term of the right-hand side of expression (5.193) is independent of the other two terms, we find

$$T_2(z) = D_2(z) \frac{A_1(V_1(z))}{V_2(z)}. \tag{5.194}$$

Substituting expressions (5.155) and (5.169) for the PRD and PRI priority queue respectively gives

$$\begin{aligned} T_2(z) =& \frac{1-\rho_{T,eff}}{\lambda_2} \left\{ \frac{(zA_1(V_1(z))-1)(A(0,V_2(z))-A_1(0))}{(V_2(z)-1)(A_1(0)z - A(0,V_2(z)))} \right. \\ &+ \frac{(zA_1(V_1(z)) - A(Y_1(V_2(z)),V_2(z)))C(z)}{A(Y_1(V_2(z)),V_2(z))} \\ &\left. \times \frac{(A_1(0)A(V_1(z),V_2(z)) - A_1(V_1(z))A(0,V_2(z)))(z-1)}{(A_1(0)z - A(0,V_2(z)))(zA_1(V_1(z)) - A(V_1(z),V_2(z)))} \right\}, \end{aligned} \tag{5.195}$$

with $C(z)$ given by (5.178) or (5.179) for the PRD or PRI case respectively.

Finally, the pgf of the waiting time of a random packet is given by

$$T(z) = \frac{\lambda_1}{\lambda_T} T_1(z) + \frac{\lambda_2}{\lambda_T} T_2(z) \tag{5.196}$$

$$= \frac{1-\rho_1}{\lambda_T} \frac{z-1}{z - A_1(S_1(z))} \frac{A_1(S_1(z))-1}{S_1(z)-1} \tag{5.197}$$

$$+\frac{1-\rho_{T,eff}}{\lambda_T}\left\{\frac{(zA_1(V_1(z))-1)(A(0,V_2(z))-A_1(0))}{(V_2(z)-1)(A_1(0)z-A(0,V_2(z)))}\right.$$
$$+\frac{(zA_1(V_1(z))-A(Y_1(V_2(z)),V_2(z)))C(z)}{A(Y_1(V_2(z)),V_2(z))}$$
$$\left.\times\frac{(A_1(0)A(V_1(z),V_2(z))-A_1(V_1(z))A(0,V_2(z)))(z-1)}{(A_1(0)z-A(0,V_2(z)))(zA_1(V_1(z))-A(V_1(z),V_2(z)))}\right\}.$$

5.8 Numerical examples

In this section, we discuss some numerical examples. Since the class-1 characteristics are the same as in the previous chapter, we solely focus on class-2 performance measures. More precisely, we focus on the comparison of the PR priority scheduling discipline - discussed in the previous chapter - and the PRD and PRI priority scheduling disciplines - analyzed in this chapter.

5.8.1 Input processes

We first briefly summarize the most important characteristics of the arrival and service processes we consider in this section.

The arrival process

The pgf of the number of per-slot class-1 and class-2 arrivals is given by

$$A(z_1,z_2)=\left(1-\frac{\lambda_1}{N}(1-z_1)-\frac{\lambda_2}{N}(1-z_2)\right)^N. \tag{5.198}$$

N is chosen 16 in all figures in this section.

The means of the total, class-1 and class-2 number of per-slot arrivals are given by λ_T, λ_1 and λ_2 respectively.

The service process

In most figures of this section, the service times of both classes are assumed deterministic

$$S_j(z)=z^{\mu_j}, \tag{5.199}$$

$j=1,2$, with μ_j the class-j service time.

In order to study the influence of the variance of the class-2 service times on the difference between the PRD and PRI case, we use class-2 service times

which are equal to $\mu_2^{(1)}$ with probability p_2 and equal to $\mu_2^{(2)}$ with probability $1 - p_2$, i.e.,

$$S_2(z) = p_2 z^{\mu_2^{(1)}} + (1 - p_2) z^{\mu_2^{(2)}}. \tag{5.200}$$

In order to study the influence of Var$[s_2]$, p_2, $\mu_2^{(1)}$ and $\mu_2^{(2)}$ will be varied so that μ_2 is kept constant and Var$[s_2]$ is varied from 0 to infinity (in discrete steps).

5.8.2 Influence of load

System contents

In Figures 5.3 and 5.4, we show the mean value and the variance of the class-2 system contents as functions of the total load, with the class-1 and class-2 service times deterministically equal to 20 and with α - $\alpha = \rho_1/\rho_T$ - equal to 0.25, 0.5 and 0.75. In both figures, we have shown the curves for the PR and the PRD and PRI priority queues. Because the service times are assumed to be deterministic the latter two scheduling types are equivalent. We furthermore show - on this figure and the other figures presented in this section - the vertical asymptotes of the curves in case of PRD and PRI, which equal the values of the *arrival* load for which the *effective* load equals 1. On the right of this asymptotes, the priority repeat queues are instable. It is seen that the mean value and the variance of the class-2 system contents can be considerably larger in case of the PRI and PRD cases. This is because of the extra load that is added because of the repeats of class-2 packets.

Figure 5.5 shows the mean class-2 system contents as a function of α, with deterministic service times of 20 slots and with ρ_T equal to 0.7, 0.8 and 0.9. We show the curves for the PR and the PRD (and PRI) priority queues. It can be seen that the influence of repeats of class-2 service times is especially high when α equals mediocre values, i.e., lies around 0.5. Indeed, for low α, almost no class-1 packets are arriving in the system and thus the interruptions of class-2 service times are scarce. For high α, the mean class-2 system contents is low simply because the arrival rate of class-2 packets is low. Note that - in the case of PRD (and PRI) and $\rho_T = 0.9$ - the system is instable for α between (approximately) 0.37 and 0.70. For values of α in this range $\rho_{T,eff} \geq 1$.

In Figure 5.6, the mean class-2 system contents in case of PRI and PRD are depicted as functions of the total load for $\alpha = 0.25, 0.5$ and 0.75 respectively. The class-1 service times are deterministically equal to 20 slots and the class-2 service time process is given by expression (5.200) with $\mu_2^{(1)} = 10$ and $\mu_2^{(2)} = 30$ in such a way that $\mu_2 = 20$ (or thus $p_2 = 0.5$ in this case). It is seen that the mean class-2 system contents may differ considerably for both priority scheduling disciplines. It is also seen that the mean class-2 system contents in the PRD priority queue is lower than the mean class-2 system contents in the

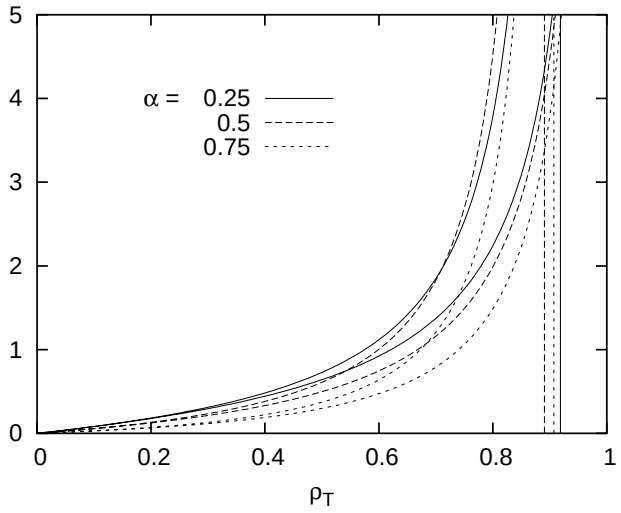


Figure 5.3: Mean class-2 system contents versus the total arrival rate for the PR (lower curves) and the PRD and PRI priority (upper curves) scheduling disciplines and with the service times deterministic ($\mu_1 = \mu_2 = 20$)

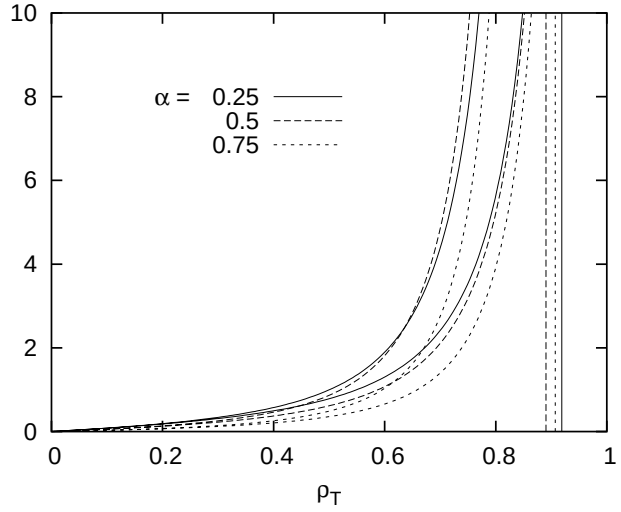


Figure 5.4: Variance of the class-2 system contents versus the total arrival rate for the PR (lower curves) and the PRD and PRI priority (upper curves) scheduling disciplines and with the service times deterministic ($\mu_1 = \mu_2 = 20$)

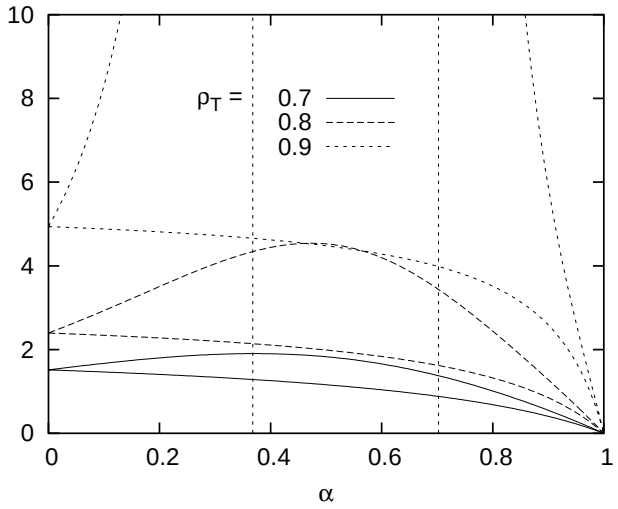


Figure 5.5: Mean class-2 system contents versus the fraction of class-1 load for the PR (lower curves) and the PRD and PRI priority (upper curves) scheduling disciplines and with the service times deterministic ($\mu_1 = \mu_2 = 20$)

PRI priority queue. This is because a long class-2 service time (30 slots) which is interrupted is resampled in the PRD priority queue and is resampled to 10 slots with a probability equal to 0.5 in the next service attempt. Obviously, a service time of 10 slots can also be resampled to one with 30 slots when interrupted, but since the probability that a 30 slots service time is interrupted is larger than that a 10 slots service time is interrupted, the resampling of the class-2 service times decreases the (mean) class-2 system contents.

Packet delay

Figure 5.7 depicts the mean class-2 packet delays as functions of the total load, with deterministic service times, $\mu_1 = 2$, $\mu_2 = 20$ and with α equal to 0.25, 0.5 and 0.75. It is seen that the mean delay of the class-2 packets is significantly higher in the case of the PRI (and PRD) priority scheduling. Again this is due to the repeats of the class-2 packets.

Furthermore, Figure 5.8 shows the mean class-2 packet delay in case of the PRD and PRI priority scheduling disciplines for $\alpha = 0.25, 0.5$ and 0.75 respectively. The class-1 service times are deterministically equal to 2 slots and the class-2 service time process is given by expression (5.200) with $\mu_2^{(1)} = 10$ and $\mu_2^{(2)} = 30$ in such a way that $\mu_2 = 20$. It is again seen that the PRD priority queue performs better - in terms of mean class-2 delays - than the PRI priority queue when the variance of the class-2 service times is larger than 0.

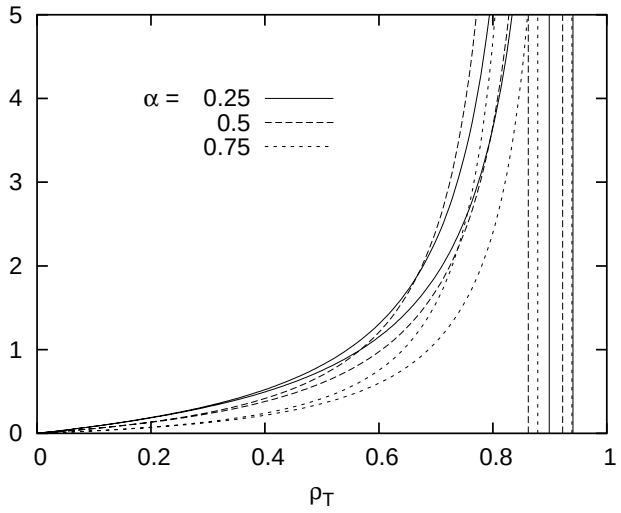


Figure 5.6: Mean class-2 system contents versus the total arrival rate for the PRD (lower curves) and the PRI priority (upper curves) scheduling disciplines and with the class-2 service times variable ($\mu_1 = \mu_2 = 20$)

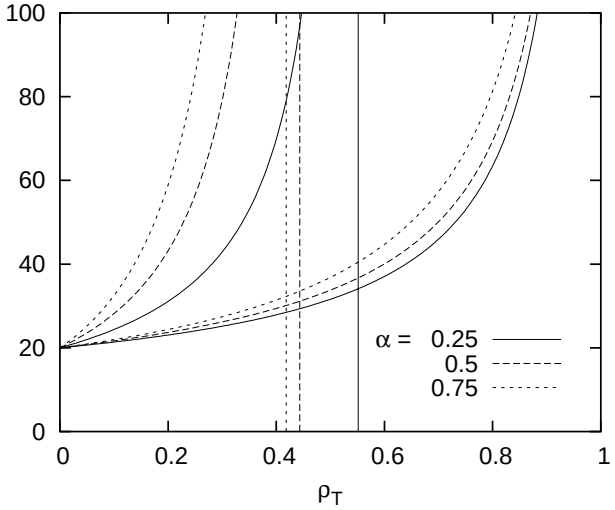


Figure 5.7: Mean class-2 packet delays versus the total arrival rate for the PR (lower curves) and the PRD and PRI priority (upper curves) scheduling disciplines and with the service times deterministic ($\mu_1 = 2, \mu_2 = 20$)

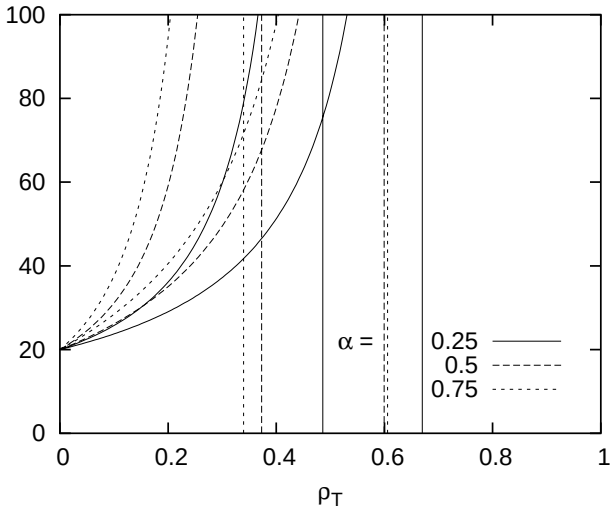


Figure 5.8: Mean class-2 packet delays versus the total arrival rate for the PRD (lower curves) and the PRI priority (upper curves) scheduling disciplines and with the class-2 service times variable ($\mu_1 = 2, \mu_2 = 20$)

5.8.3 Influence of service times

System contents

Figure 5.9 shows the mean class-2 system contents in case of PR and PRD (and PRI) as functions of the class-1 service time μ_1, when class-2 service times equal 20 slots, the total arrival load is 0.75 and for α equal to 0.25, 0.5 and 0.75. The service times of both classes are deterministic. In case of PR scheduling, the mean system contents increase with increasing μ_1. In case of the PRI and PRD scheduling, two counter-acting effects can be observed: firstly, longer class-1 packets increase the build-up periods for class-2 packets in the queue (i.e., longer periods when the server is busy with class-1 packets), thereby increasing the mean class-2 packet delay. Secondly, longer class-1 packets (while keeping the class-1 arrival load constant) means less class-1 packet arrivals, thus decreasing the probability of a class-2 packet's service getting preempted and having to be repeated. This has a decreasing effect on the mean class-2 system contents. The latter effect is important for small class-1 service times. Indeed, it can be seen on the figure that for low μ_1 the mean class-2 system contents increase dramatically with decreasing μ_1. So, in the case of the repeat scheduling disciplines there is an optimum for μ_1 for which the mean class-2 system contents becomes minimal.

Figure 5.10 depicts the mean class-2 system contents in the PR and PRD (and PRI) priority queues as functions of the (mean) class-2 service times, for $\mu_1 =$

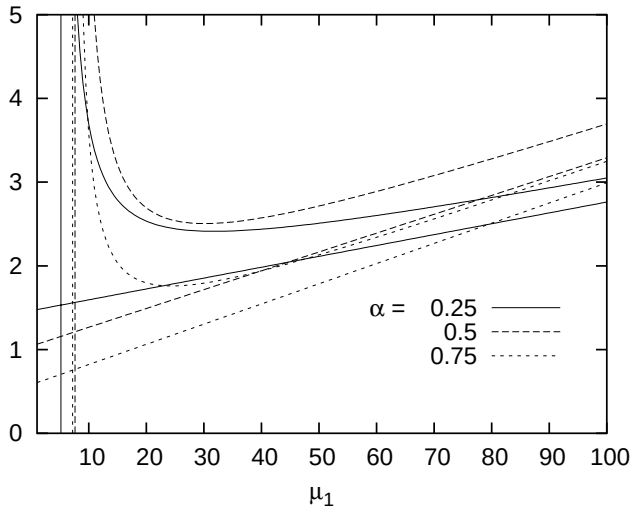


Figure 5.9: Mean class-2 system contents versus the class-1 service time for both the PR (lower curves) and the PRI and PRD priority (upper curves) scheduling disciplines and with the service times deterministic ($\rho_T = 0.75, \mu_2 = 20$)

20, $\rho_T = 0.75$ and the different values of α. We see that for low μ_2, the mean class-2 system contents in both priority queues are high. This is due to the fact that λ_2 is relatively large for low μ_2 (in order to keep ρ_2 constant) and thus many class-2 packets arrive to - and have to be stored in - the system. For increasing μ_2, the mean system contents in case of PR decreases, while in case of PRI (and PRD) the mean system contents first decreases and then increases again. The latter is due to the fact that service of longer class-2 packets have a higher probability of being preempted by arriving class-1 packets. For a particular μ_2 these numbers of interruptions and repeats are too large to still have a stable system.

Figure 5.11 shows the mean class-2 system contents in the PR, PRD and PRI priority queues as functions of the variance of the class-2 service times. The total load is fixed at 0.75, the class-1 service times are deterministically equal to 20 slots and $\alpha = 0.25$. The class-2 service time process is characterized by expression (5.200), with $\mu_2^{(1)}$ equal to 1, $\mu_2^{(2)}$ varying and p_2 also varying but such that μ_2 is kept constant (equal to 20 slots). This figure clearly shows that the mean class-2 system contents can be very different in the PRD and PRI priority queues respectively. Indeed, when the variance is large, long service times have a high probability of being resampled in service times of 1 slot in the PRD case, while this is not the case in a PRI priority queue. Also notice that the mean class-2 system contents is lower in the PRD case than in the PR case for high variances. Indeed, since long service times can be resampled to

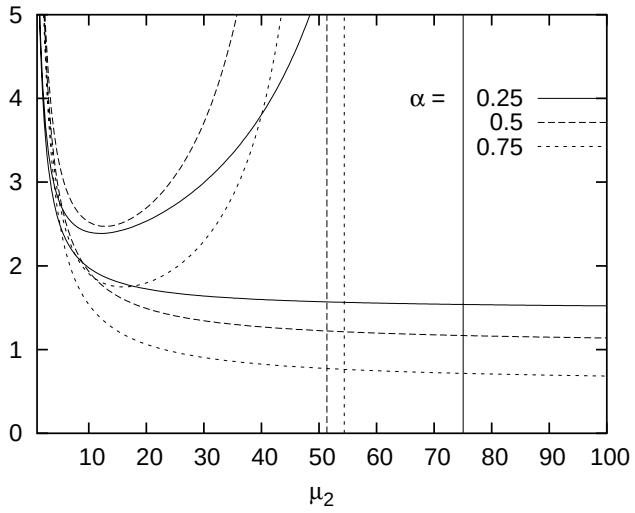


Figure 5.10: Mean class-2 system contents versus the class-2 service time for both the PR (lower curves) and the PRI and PRD priority (upper curves) scheduling disciplines and with the service times deterministic ($\rho_T = 0.75, \mu_1 = 20$)

service times of 1 slot, the resampled version of the complete service time can be smaller than the residual part of the original service time.

Packet delay

Similar conclusions can be drawn for the mean class-2 packet delay from Figures 5.12 and 5.13. In Figure 5.12 the same input parameters are chosen as in Figure 5.9, while the input parameters in Figure 5.13 are identical to the ones in Figure 5.11. It is seen that the PRI and PRD scheduling types have an optimum μ_1 (for which the mean class-2 packet delay is minimal) and that the mean delay differs considerably in the PRD and PRI priority queues respectively if the variance of the class-2 service times is large.

5.9 Concluding remarks

In this chapter, we analyzed two types of preemptive repeat priority queues, namely a PRD and PRI priority queue. In both cases, we defined supplementary variables - besides the system contents of both priority classes - in order to construct a Markov-chain. In the PRD case, defining the remaining service time of the packet in service was enough to construct a Markov-chain. In the PRI case, we needed to keep track of the complete service time of the oldest

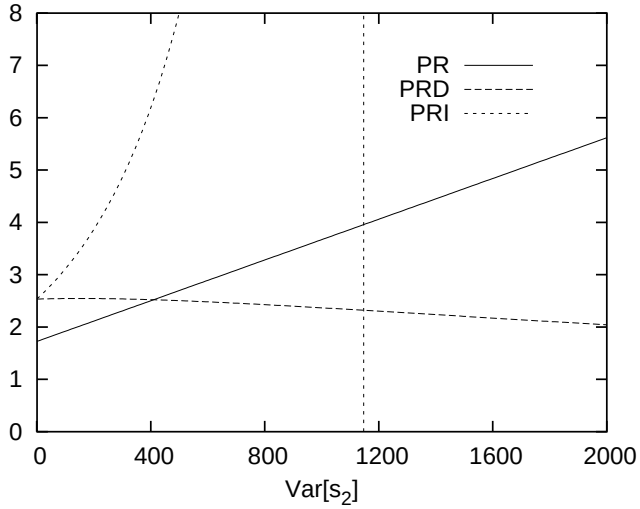


Figure 5.11: Mean class-2 system contents versus the variance of the class-2 service times for the PR, PRD and PRI case ($\rho_T = 0.75, \mu_1 = \mu_2 = 20$)

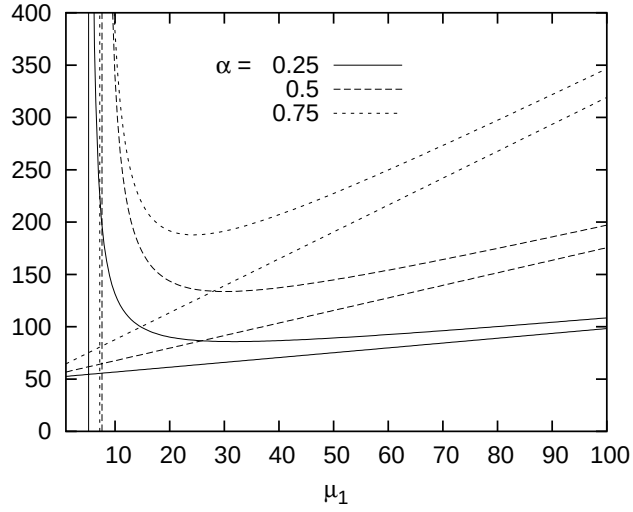


Figure 5.12: Mean class-2 packet delays versus the class-1 service time for both the PR (lower curves) and the PRI and PRD priority (upper curves) scheduling disciplines and with the service times deterministic ($\rho_T = 0.75, \mu_2 = 20$)

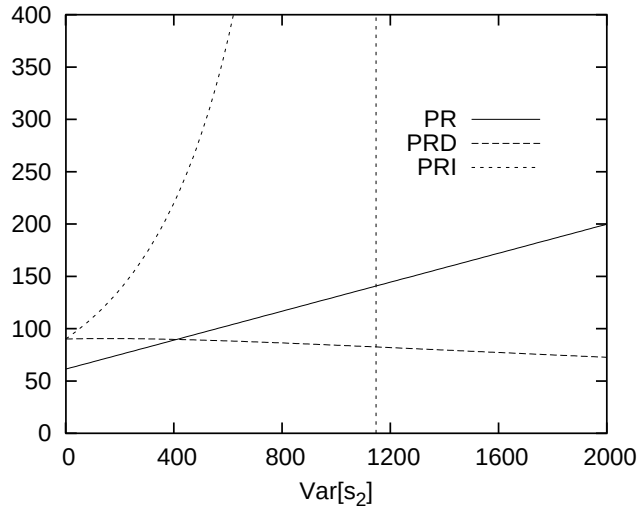


Figure 5.13: Mean value of class-2 packet delays versus the variance of the class-2 service times for the PR, PRD and PRI case ($\rho_T = 0.75, \mu_1 = \mu_2 = 20$)

class-2 packet as well. We thus had to calculate a three-dimensional pgf in the PRD case and a four-dimensional pgf in the PRI case. These pgf's were the starting points of all calculations of the performance measures of these priority queues.

Note that an infinite sum is part of the solution in the class-2 pgf's (system contents, packet delay, ...) of the PRI case. This infinite sum does not give problems when evaluating the means and variances of these variables, but the calculation of higher moments could turn out to be of a more complex nature in practice - although it is feasible in theory. Furthermore, the infinite sum prohibits us - at this time? - to find approximate tail probabilities of the class-2 variables.

We compared the performance measures of a PR priority queue - analyzed in the previous chapter - of a PRD priority queue and of a PRI priority queue. Main conclusions are firstly the fact that repeats of class-2 service times in the PRI and PRD priority queues may considerably deteriorate the performance of these systems, especially when the arrival rate of class-1 packets is large. Secondly, the results of the PRD priority queue and the PRI priority queue are only similar when the variance of the class-2 service times is sufficiently small. Both scheduling disciplines are only identical when this variance is zero.

Chapter 6

Conclusions

In this last chapter, we summarize the main results of this dissertation and briefly describe some possible extensions and related topics.

6.1 Summary

In this dissertation, we comprehensively described the analysis of *discrete-time* queues with different types of *priority* scheduling disciplines. We assumed two priority classes throughout the dissertation. The arrival process was assumed to be i.i.d. from slot-to-slot, but the number of per-slot arrivals of both priority classes may be correlated, i.e., so-called *structured input*. We started by analyzing a priority queue with service times of one slot and extended this to general service times for both classes. The distributions of the service times could furthermore be different for both classes.

In chapter 2, we analyzed a priority queue with single-slot service times. This is a fairly easy model and analysis which was mainly used to familiarize the reader with pgf's, the way in which the performance measures of interest are extracted from the obtained pgf's and the specific difficulties introduced by the priority scheduling discipline. Furthermore, this type of service times is useful in practice, e.g. in telecommunication networks where the packets floating through the network are all of the same size (e.g. in ATM).

In chapters 3, 4 and 5, we extended this inital model to a model with generally distributed service times. In case of general service times, 3 main priority scheduling disciplines are distinguished in the literature and analyzed in this dissertation. The first one is the *non-preemptive* priority scheduling discipline. In this discipline, service times are never interrupted, i.e., once a unit starts service it stays in the server until its service is completed. This type of priority queue was analyzed in chapter 3. The second type is the *preemptive resume*

priority scheduling discipline, studied in chapter 4. In this discipline, a new arrival of a unit of higher priority than the one served during that slot interrupts this latter unit's service at the end of the slot. When all higher priority units have left the system the interrupted unit resumes its service, i.e., the not-yet-served part of this unit has to be served after the interruption. The last scheduling type is the *preemptive repeat* priority scheduling discipline, which we analyzed in section 5. This type of priority scheduling differs from the preemptive resume priority in the fact that a unit whose service was interrupted by higher priority arrivals repeats its service from the start. Thus the unit's service time has to be repeated completely.

Throughout the different queueing models, we used a fairly general analysis method. Firstly, a Markovian description of the system was found. This Markovian description exists out of a number of stochastic variables. We adopted the probability generating functions (pgf's) technique in order to analyze these priority queues. The joint pgf of the steady-state versions of the random variables defined in the Markovian description is calculated. From this joint pgf all further pgf's and performance measures of interest are calculated (in steady-state). These pgf's are: the joint pgf and the marginal pgf's of the number of units of both classes in the system at the beginning of a random slot, the marginal pgf of the total number of units in the system, the joint pgf of the number of units of both classes in the queue (i.e., without the unit in service), the joint pgf of the unfinished work of both classes, the pgf's of the delays of units of both classes, the pgf of the delay of a random unit and finally the pgf's of the waiting times of units of both classes. From these pgf's, it was shown how moments and approximate tail probabilities of the stochastic variables of interest can be calculated.

6.2 Possible extensions and related topics

As with all research, this dissertation - although described as a self-containing unit - has many possible extensions. To conclude this dissertation, we describe some of these extensions and other related topics and refer the interested reader to related work.

The most direct extension is the extension of the input model of the priority queue. Firstly, we have only described the analysis of a *two-class* priority queue in this dissertation. This could be extended to more classes (and even a general number of classes). For example in the case of the non-preemptive priority scheduling, we have already analyzed a three-class priority queue in [Walraevens et al. 2003b] and a priority queue with a general number of classes in [Walraevens et al. 2004c]. Secondly, the arrival process could be extended to a correlated arrival process, i.e., in which the numbers of arrivals during a slot depend on the numbers of arrivals of previous slots. The easiest extension is a model in which the number of arrivals during one slot only depends on the number of arrivals during the previous slot, but other extensions

are also possible (e.g. the use of so-called 'train arrivals', see e.g. [Wittevrongel 1998] for a thorough discussion and analysis of single-class discrete-time FIFO buffers with this type of arrival process). Finally, the service time process could be extended in a couple of ways. In this dissertation, we have only studied single-server queueing models. This could be extended to multi-server models. This is e.g. already done by Gao et al. [2004] in the case of a preemptive resume priority scheduling discipline and geometric service times. Furthermore, service interruptions or vacations could also be incorporated in the queueing model. Examples of studies of priority queues with vacations are studied by Blondia [1987, 1989], Sandhu and Posner [1989], Katsaros and Langaris [1995] and Langaris and Katsaros [1997].

In this dissertation, we studied the system characteristics and mainly focussed on the system contents and delay. Another interesting topic is studying the *output process* of priority queues. The study of this output process is important when studying cascades of queues, since the output process of one queue is (related to) the input process of another. For example in [Stanford 1991, 1997, He and Stanford 1999], the departure process of priority queues is studied in more detail. In our model, we could e.g. extend the supplementary variable technique in order to keep track of the length of the ongoing busy period (of a certain class).

We have analyzed the *steady-state* performance characteristics in this dissertation. Studying the transient behavior could be interesting as well though. This could be done by combining our methods and the method used in Bruneel [1991], where the transient behavior of a single-class discrete-time queue with service times of one slot was analyzed.

Another extension is combining the different priority scheduling disciplines analyzed in this dissertation, i.e., combining the NP, the PR, the PRD and/or the PRI priority scheduling disciplines. This could e.g. be done by splitting the service times of the low-priority units in several parts and using a different priority scheduling discipline for each part. Hokstad [1978], Sandhu and Posner [1989], Cho and Un [1993], Paterok and Ettl [1994] and Machihara [1995] propose and analyze combinations of NP and PR priority scheduling disciplines. Queues with a priority scheduling discipline which is some kind of mixture between the NP and the PRI or PRD cases are analyzed by Adiri and Domb [1984], Cho and Un [1993]. Yoon and Un [1991], Drekic and Stanford [2001], Drekic and Grassmann [2002] and Drekic [2003] investigate combinations between PR and PRI priority scheduling disciplines. Finally, Hong and Takagi [1997] study a queue with a priority scheduling which is a combination of NP, PR and PRI.

In the remainder of this section, we will briefly touch upon analyses of some related (priority) queues. Priority queueing *networks* are studied by Peterson [1991], Chen and Zhang [1998, 2000], Afèche [2003] and Kouvatsos and Awan [2003]. Tijms [1974] investigates a control policy for a priority queue with removable servers. This policy turns the server off when the system is empty

and turns it back on when a given linear combination of the numbers of units of all priority classes in the system exceeds a given value. Schaack and Larson [1986] study a multi-server non-preemptive cut-off priority queue. In this type of priority queues, the system deliberately queues arriving lower priority units whenever the number of busy servers exceeds a given priority-dependent number, instead of serving them directly. This is done to keep some of the servers idle to be able to serve arriving higher priority units immediately. A priority queue with set-up times is studied by Takagi [1990]. In this system, a set-up time is required before initiating service from the system's idle state. Sidi [1987, 1988] analyzes a discrete-time priority queue with partial interference. More precisely, a number of traffic classes are served according to a priority scheduling discipline, while one class uses a random access scheme (i.e., a cell of this class (if any) is tried to be served with an assigned probability in [Sidi 1987], while there is some kind of correlation between the probabilty of attempting to be served in consecutive slots in [Sidi 1988]). If a cell of one of the priority classes and a cell of this latter class are attempted to be served simultaneously, there is no service at all. Choi et al. [1998b] analyze a priority queue with random order of service within each priority class, while van der Mei et al. [2003] study a priority processor sharing queue. Antal and Bíró [2000] study a priority system with one high-priority source and a number of low-priority sources. The time-axis is divided into frames (which consist of a fixed number of slots). High-priority cells can be transmitted in every slot, while the low-priority sources can use only one slot in every frame (evidently, a cell can only be transmitted in that slot if no high-priority cells are present). Lee and Choi [2001] study a queueing system with a priority scheduling discipline and a push-out scheme. Leemans [2001] analyzes a two-server two-class priority queue. The priority order is different for the two servers: in the first server one of the classes has priority over the other, while it is the other way around in the other server. Choi and Park [1990], Falin et al. [1993], Takahashi et al. [1999], Artalejo et al. [2001] and Gómez-Corral [2002] study retrial queues with an NP priority scheduling discipline. In [Ozawa 1992], an NP priority queue with gates is analyzed. Chao [1994] analyzes a priority tandem queue. Mandjes [2003] investigates pricing strategies in networks with priority. Finally, Maertens et al. [2004] analyze a priority queue with priority jumps. In this model, units can jump to a higher priority queue while waiting.

Appendix A

The function $Y_1(z)$

In this appendix, we describe more fundamental details on the function $Y_1(z)$ (or the function $Y(z)$ of chapter 2) which is an important function in this dissertation. First, we describe Rouché's theorem and its use in finding the solution $z_1 = Y_1(z_2)$ of $z_1 - S_1(A(z_1, z_2)) = 0$, for $|z_1| < 1$ and $|z_2| < 1$. Furthermore, we explain the behavior of $Y_1(z)$ on the real axis (outside the unit disk). For ease of notation, define $E_1(z_1, z_2) = S_1(A(z_1, z_2))$, with $A(z_1, z_2)$ the two-dimensional pgf of the numbers of class-1 and class-2 arrivals in a random slot and $S_1(z)$ the pgf of the service time of a random class-1 unit. Furthermore, we define $E_1^{(1)}(z_1, z_2)$ as the first derivative of $E_1(z_1, z_2)$ in z_1. Note that $E_1(z_1, z_2)$ is also a two-dimensional pgf, more precisely, it is the joint pgf of the numbers of class-1 and class-2 arrivals during the service time of a class-1 packet.

A.1 Rouché's theorem

Theorem A.1 *Let $f(z)$ and $g(z)$ be two analytic functions inside and on a closed contour C in the complex z-plane such that $|g(z)| < |f(z)|$ for all z on C. Then the functions $f(z)$ and $f(z) + g(z)$ have the same number of zeros inside C.*

A.2 Determination of $Y_1(z)$ for $|z| < 1$

We use Rouché's theorem to prove that for each z_2 (with $|z_2| < 1$), there exists a unique solution of $z_1 - E_1(z_1, z_2) = 0$ for z_1 in the unit disk.

First, we look for a contour C to apply Rouché's theorem. Therefore, we first compare the functions $\hat{f}(|z_1|) \triangleq |z_1|$ and $\hat{g}(|z_1|) \triangleq E_1(|z_1|, |z_2|)$, with z_2 a fixed point inside the unit disk and $|z_1|$ in the range $[0, 1]$. Firstly, it is easily seen that

$\hat{f}(|z_1|)$ and $\hat{g}(|z_1|)$ are continuously increasing functions in $[0, 1[$. Furthermore, since

$$\begin{aligned}\hat{g}(0) =& E_1(0, |z_2|) && \text{(A.1)}\\ \geq& E_1(0, 0) && \text{(A.2)}\\ >& 0 && \text{(A.3)}\\ =& \hat{f}(0), && \text{(A.4)}\end{aligned}$$

and since

$$\begin{aligned}\hat{g}(1) =& E_1(1, |z_2|) && \text{(A.5)}\\ <& 1 && \text{(A.6)}\\ =& \hat{f}(1), && \text{(A.7)}\end{aligned}$$

$\hat{f}(|z_1|) - \hat{g}(|z_1|)$ has exactly one zero z_* in $[0, 1[$. In the following we will define the contour C as the circle with radius R with R a randomly chosen real number in the interval $]z_*, 1[$. Note that it directly follows from the choice of R that

$$\hat{g}(R) < \hat{f}(R). \tag{A.8}$$

Now, define

$$\begin{aligned}f(z_1) \triangleq& z_1 && \text{(A.9)}\\ g(z_1) \triangleq& E_1(z_1, z_2), && \text{(A.10)}\end{aligned}$$

with z_2 still a fixed point inside the unit disk. $f(z_1)$ is analytic in the whole complex plane and since $E_1(z_1, z_2)$ is a pgf, $g(z_1)$ is analytic inside the unit circle. Both $f(z_1)$ and $g(z_1)$ are thus analytic inside and on the contour C - a circle inside the unit disk. Since

$$|f(z_1)| = \hat{f}(R), \tag{A.11}$$

and - keeping in mind that $|z_2| < 1$ -

$$\begin{aligned}|g(z_1)| =& |E_1(z_1, z_2)| && \text{(A.12)}\\ \leq& E_1(|z_1|, |z_2|) && \text{(A.13)}\\ =& \hat{g}(R), && \text{(A.14)}\end{aligned}$$

on the contour C ($|z_1| = R$), it follows from (A.8) that $|g(z_1)| < |f(z_1)|$ on C. It then follows from Rouché's theorem that $z_1 - E_1(z_1, z_2)$ has the same number of zeros as z_1 inside the contour C. The latter has one solution - namely $z_1 = 0$

- and thus $z_1 - E_1(z_1, z_2)$ has a unique solution for z_1 inside C. Since the contour C can be chosen arbitrarily close to the unit disk - inequality (A.8) is valid for every $R \in]z_*, 1[$ - $z_1 - E_1(z_1, z_2)$ has one solution inside the unit disk which is denoted by $Y_1(z_2)$ (for each z_2, with $|z_2| < 1$).

Note that another possible choice of C is the unit circle. However, since pgf's are not necessarily analytic on the unit circle this would have restricted the choice of the pgf $E_1(z_1, z_2)$ in order for Rouché's theorem to work, in contrast with the proof given in this appendix, which is valid for every pgf $E_1(z_1, z_2)$.

The next question is whether $Y_1(z)$ itself is an analytic function inside the unit circle (i.e., for $|z| < 1$)? Since $Y_1(z)$ is a pgf - in the dissertation a stochastic variable is found with a pgf equal to $Y_1(z)$ - $Y_1(z)$ is indeed an analytic function inside the unit circle. Furthermore we have $Y_1(1) = 1$.

A.3 Outside the unit disk

To calculate the (tail) probabilities of the stochastic variable corresponding with the pgf $Y_1(z)$, the behavior of $Y_1(z)$ outside the unite disk is important, and more precisely the singularities of this function. To calculate the tail probabilities only the dominant singularity, lying on the positive real axis, is important. Therefore, we limit the function $Y_1(z)$ for values of z on the positive real axis in the remainder. First, we give a useful theorem, the *implicit function theorem*. There are many versions of this theorem, depending on the domains and types of functions one is dealing with. The following version is suitable for our case.

A.3.1 The implicit function theorem

Theorem A.2 *Given a real function $F(z_1, z_2)$, which is continuously differentiable in the neighborhood of $z_1 = z_1^{(0)}, z_2 = z_2^{(0)}$. If $F(z_1^{(0)}, z_2^{(0)}) = 0$ and $F^{(1)}(z_1^{(0)}, z_2^{(0)}) \neq 0$ then there exists a unique continuously differentiable function $f(z_2)$ in the neighborhood of $z_2^{(0)}$ with $z_1^{(0)} = f(z_2^{(0)})$ and $F(f(z_2), z_2) = 0$.*

A.3.2 $Y_1(z)$ on the positive real axis

We have already proved that $Y_1(z)$ is continuously differentiable in the interval $[0, 1[$. Can we now extend the function $Y_1(z)$ for higher z? Starting in the point $z = 1, Y_1(z) = 1$, we use the implicit function theorem to investigate in what region $Y_1(z)$ is continuously differentiable. Denoting $F(z_1, z_2) = z_1 - E_1(z_1, z_2)$, it can be seen that the implicit function theorem will work as long as $F(z_1, z_2)$ is continuously differentiable and $F^{(1)}(Y_1(z), z) \neq 0$. $F(z_1, z_2)$

is continuously differentiable iff $E_1(z_1, z_2)$ is continuously differentiable and $F^{(1)}(Y_1(z), z) \neq 0$ iff

$$E^{(1)}(Y_1(z), z) \neq 1. \tag{A.15}$$

We start at $z = 1$. We assume that $E_1(z_1, z_2)$ is continuously differentiable in $(1, 1)$. Furthermore, condition (A.15) is true, since $E^{(1)}(Y_1(1), 1) = \rho_1 < 1$. We furthermore assume that $E^{(1)}(Y_1(z), z)$ is a strictly increasing function in z inside the region of convergence of $E(Y(z), z)$ (which is true for most 'normal' arrival and service processes). So, by increasing z, $E^{(1)}(Y_1(z), z)$ will get closer to 1. As long as it does not reach 1, $Y_1(z)$ can be defined for that particular z. For a specific z however, denoted by z_B, $E^{(1)}(Y_1(z), z)$ becomes 1, and the implicit function theorem can no longer be used. So $Y_1(z)$ can be defined in the interval $[0, z_B[$.

The last question remaining is what happens for $z = z_B$? Note that it cannot be proved through the implicit function theorem that $Y_1(z_B)$ exists (or, in other words, stays finite). We can however illustrate this by noting that $z_1 - E_1(z_1, z_2)$ has 2 solutions for $z_2 \in [0, z_B[$. This can for instance be seen in Figure A.1. In this figure, we have shown z_1, and $E_1(z_1, z_2)$ for four z_2's, denoted by $z_2^{(i)}$ ($i = 1, \ldots, 4$), with

$$z_2^{(1)} < z_2^{(2)} < z_2^{(3)} = z_B < z_2^{(4)}, \tag{A.16}$$

with $E_1(z_1, z_2)$ a two-dimensional pgf (this is just an example; the specific expression of the pgf $E_1(z_1, z_2)$ is not important). We see that $z_1 - E_1(z_1, z_2)$ has two solutions if $z_2 < z_B$. The smallest solution is $Y_1(z_2)$ and we denote the other solution by $Y_1^*(z_2)$. Notice that for $z_2 < z_B$ this second solution always exists, since $E^{(1)}(Y_1(z_2), z_2) < 1$ and thus $E_1(z_1, z_2)$ and z_1 will intersect in this point. Since we assumed throughout this dissertation that $E_1(z_1, z_2)$ and $E^{(1)}(z_1, z_2)$ go to infinity for z_1 equal to the convergence radius of $E_1(z_1, z_2)$ (for the specific z_2), $E_1(z_1, z_2)$ and z_1 intersect once more, resulting in the second solution $Y_1^*(z_2)$. It can be proven that $Y_1^*(z)$ is a continuously differentiable function for $z < z_B$ in a similar way as this is proven for $Y_1(z)$. Since $E_1(z_1, z_2)$ is a strictly increasing function in z_2 it can be seen that $Y_1(z_2)$ is a strictly increasing function while $Y_1^*(z_2)$ is a strictly decreasing function (see Figure A.1: $Y_1(z_2^{(1)}) < Y_1(z_2^{(2)})$ and $Y_1^*(z_2^{(1)}) > Y_1^*(z_2^{(2)})$). For z_2 going to z_B, $Y_1(z_2)$ and $Y_1^*(z_2)$ grow nearer to each-other, resulting in $Y_1(z_B) = Y_1^*(z_B)$. It can thus also be seen that $Y_1(z_B)$ will always be finite, since for all $z_* < z_B$, $Y_1(z_*) < Y_1(z_B) = Y_1^*(z_B) < Y_1^*(z_*)$. For $z_2 > z_B$, $E_1(z_1, z_2) > z_1$ for all z_1 resulting in the fact that $z_1 - E_1(z_1, z_2) = 0$ has no solution for $z_2 > z_B$ (see the curve of $E_1(z_1, z_2^{(4)})$ in Figure A.1). For the same input functions as used in Figure A.1, $Y_1(z)$ and $Y_1^*(z)$ are shown in Figure A.2.

So, summarizing, $z_1 - E_1(z_1, z_2) = 0$ has 2 solutions for z_1 when $z_2 \in [0, z_B[$, which coincide in z_B. For $z_2 > z_B$, $z_1 - E_1(z_1, z_2) = 0$ has no solution. $Y_1(z)$

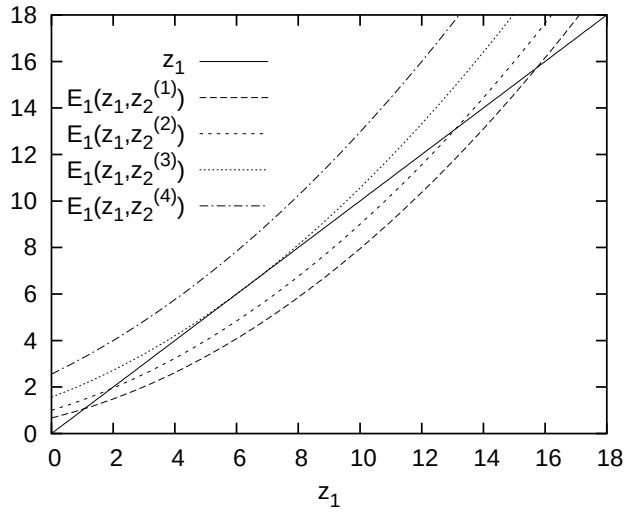


Figure A.1: z_1 and $E_1(z_1, z_2)$ for different z_2

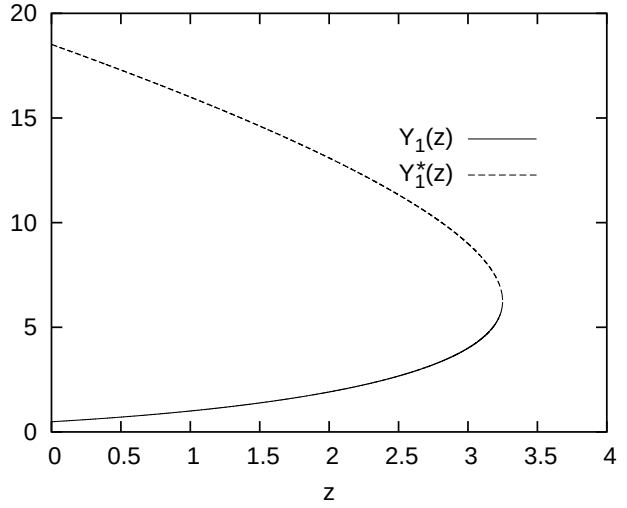


Figure A.2: $Y_1(z)$ and $Y_1^*(z)$

thus has a branch point z_B on the real positive axis and a branch cut which is chosen on the positive real axis (starting in this branch point). This branch point is furthermore the dominant singularity of $Y_1(z)$.

Bibliography

Abate, J. and Whitt, W. (1997). Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems*, 25(1-4):173–233.

Adiri, I. and Domb, I. (1984). Mixing of non-preemptive and preemptive repeat priority disciplines. *European Journal of Operational Research*, 18(1):86–97.

Afèche, P. (2003). Delay performance in stochastic processing networks with priority service. *Operation Research Letters*, 31(5):390–400.

Altinkemer, K., Bose, I., and Pal, R. (1998). Average waiting time of customers in an M/D/k queue with nonpreemptive priorities. *Computers and Operations Research*, 25(4):317–328.

Antal, C. and Bíró, J. (2000). *Analysis of a time division multiplexing method with priorities*, chapter 11, pages 165–175. System Performance Evaluation: methodologies and Applications. CRC Press.

Artalejo, J., Dudin, A., and Klimenok, V. (2001). Stationary analysis of a retrial queue with preemptive repeated attempts. *Operation Research Letters*, 28(4):173–180.

Bender, E. (1974). Asymptotic methods in enumeration. *SIAM Review*, 16(4):485–515.

Berger, A. and Whitt, W. (2000). Workload bounds in fluid models with priorities. *Performance Evaluation*, 41(4):249–267.

Blondia, C. (1987). An M/G/1 finite capacity queue with vacations and priorities. In *Proceedings of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation (Performance '87)*, pages 305–325, Brussels.

Blondia, C. (1989). A finite capacity multi-queueing system with priorities and with repeated server vacations. *Queueing Systems*, 5(4):313–330.

Bose, I. and Pal, R. (2002). Average waiting time of customers in a priority M/D/k queue with finite buffers. *Computers and Operations Research*, 29(4):327–339.

Boxma, O., Cohen, J., and Deng, Q. (1999). Heavy-traffic analysis of the M/G/1 queue with priority classes. In *Proceedings of ITC 16*, pages 1157–1167, Edinburgh.

Brandwajn, A. (1982). A finite difference equations approach to a priority queue. *Operations Research*, 30(1):74–81.

Bruneel, H. (1983). Analysis of buffer behaviour for an integrated voice-data system. *Electronics Letters*, 19(2):72–74.

Bruneel, H. (1991). Exact derivation of transient behavior for buffers with random output interruptions. *Computers Networks and ISDN Systems*, 22:277–285.

Bruneel, H. (1993). Performance of discrete-time queueing systems. *Computers and Operations Research*, 20(3):303–320.

Bruneel, H. and Kim, B. (1993). *Discrete-time models for communication systems including ATM*. Kluwer Academic Publisher, Boston.

Bruneel, H., Steyaert, B., Desmet, E., and Petit, G. (1994). Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. *European Journal of Operational Research*, 76(3):563–572.

Buzen, J. and Bondi, A. (1983). The response times of priority classes under preemptive resume in M/M/m queues. *Operations Research*, 31(3):456–465.

Chang, J. and Harn, Y. (1992). A discrete-time priority queue with two-class customers and bulk services. *Queueing Systems*, 10:185–212.

Chao, C. (1994). A priority tandem queue with no intermediate buffer. *Journal of the Operational Research Societys*, 45(3):321–329.

Chaudhry, M. and Templeton, J. (1983). *A first course in bulk queues*. John Wiley & Sons.

Chen, H. and Zhang, H. (1998). Diffusion approximation for Kumar-Seidman network under a priority service discipline. *Operations Research Letters*, 23(3-5):171–181.

Chen, H. and Zhang, H. (2000). Stability of multiclass queueing networks under priority service disiplines. *Operations Research*, 48(1):26–37.

Chen, J. and Guérin, R. (1991). Performance study of an input queueing packet switch with two priority classes. *IEEE Transactions on Communications*, 39(1):117–126.

Cho, Y. and Un, C. (1993). Analysis of the M/G/1 queue under a combined preemptive/nonpreemptive priority discipline. *IEEE Transactions on Communications*, 41(1):132–141.

Choi, B., Choi, D., Lee, Y., and Sung, D. (1998a). Priority queueing system with fixed-length packet-train arrivals. *IEE Proceedings-Communications*, 145(5):331–336.

Choi, B., Lee, Y., and Choi, D. (1998b). Geo^{X_1},Geo^{X_2}/D/c HOL priority queueing system with random order selction within each priority class. *Probability in the Engineering and Informational Sciences*, 12(1):125–139.

Choi, B. and Park, K. (1990). The M/G/1 retrial queue with Bernoulli schedule. *Queueing Systems*, 7(2):219–228.

Choi, J., Lee, K., and Un, C. (1997). Performance comparison of nonpreemptive and preemptive priority queueing strategies in ATM packet switch with input buffers. *Performance Evaluation*, 29(3):177–194.

Cidon, I. and Sidi, M. (1990). Recursive computation of steady-state probabilities in priority queues. *Operations Reasearch Letters*, 9(4):249–256.

Cobham, A. (1954). Priority assignment in waiting line problems. *Journal of the American Operations Research Society*, 2(1):70–76.

Cox, D. (1955). The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society*, 51:433–441.

De Prycker, M. (1991). *Asynchronous transfer mode: solution for broadband ISDN*. Ellis Horwood Limited, New York.

Drekic, S. (2003). A preemptive resume queue with an expiry time for retained service. *Performance Evaluation*, 54(1):59–74.

Drekic, S. and Grassmann, W. (2002). An eigenvalue approach to analyzing a finite source priority queueing model. *Annals of Operations Research*, 112(1-4):139–152.

Drekic, S. and Stafford, J. (2002). Symbolic computation of moments in priority queues. *INFORMS Journal on Computing*, 14(3):261–277.

Drekic, S. and Stanford, D. (2001). Reducing delay in preemptive repeat priority queues. *Operations Research*, 49(1):145–156.

Drmota, M. (1997). Systems of functional equations. *Random Structures & Algorithms*, 10(1-2):103–124.

Falin, G., Artalejo, J., and Martin, M. (1993). On the single-server retrial queue with priority customers. *Queueing Systems*, 14(3-4):439–455.

Fiems, D. (2004). *Analysis of discrete-time queueing systems with vacations*. PhD thesis, Ghent University.

Fiems, D. and Bruneel, H. (2002). A note on the discretization of Little's result. *Operations Research Letters*, 30(1):17–18.

Fiems, D., Steyaert, B., and Bruneel, H. (2004). Discrete-time queues with generally distributed service times and renewal-type server interruptions. *Performance Evaluation*, 55(3-4):277–298.

Flajolet, P. and Odlyzko, A. (1990). Singularity analysis of generating functions. *SIAM Journal on discrete mathematics*, 3(2):216–240.

Gail, H., Hantler, S., and Taylor, B. (1992). On a preemptive Markovian queue with multiple servers and two priority classes. *Mathematics of Operations Research*, 17(2):365–391.

Gao, P., Wittevrongel, S., and Bruneel, H. (2004). Analysis of buffer behavior for a discrete-time multiserver preemptive priority queue with geometric service times. In *Abstracts of the Twelfth INFORMS/APS Conference*, Beijing.

Goldberg, H. (1981). Computation of state probabilities for M/M/s priority queues with customer classes having different service rates. *INFOR*, 19(1):48–58.

Gómez-Corral, A. (2002). Analysis of a single-server retrial queue with quasi-random input and nonpreemptive priority. *Computers and Mathematics with Applications*, 43(6-7):767–782.

Hadorn, D. (2000). Setting priorities for waiting lists: defining our terms. *Canadian Medical Association Journal*, 163(7):857–860.

Hashida, O. and Takahashi, Y. (1991). A discrete-time priority queue with switched batch bernoulli process inputs and constant service time. In *Proceedings of ITC 13*, pages 521–526, Copenhagen.

He, Q. and Stanford, D. (1999). Distributions of the interdeparture times in fcfs and nonpreemptive priority MMAP[2]/G[2]/1 queues. *Performance Evaluation*, 38(2):85–103.

Hokstad, P. (1978). A M/G/1 priority queue. *INFOR*, 16(2):158–170.

Hong, S. and Takagi, H. (1997). Analysis of transmission delay for a structured-priority packet-switching system. *Computer Networks and ISDN Systems*, 29(6):701–715.

Iida, K., Takine, T., Sunahara, H., and Oie, Y. (2001). Delay analysis for cbr traffic under static-priority scheduling. *IEEE/ACM Transactions on Networking*, 9(2):117–185.

Isotupa, K. and Stanford, D. (2002). An infinite-phase quasi-birth-and-death model for the non-preemptive priority M/PH/1 queue. *Stochastic Models*, 18(3):387–424.

Kao, E. and Narayanan, K. (1991). Modeling a multiprocessor system with preemptive priorities. *Management Science*, 37(2):185–197.

Kao, E. and Wilson, S. (1999). Analysis of nonpreemptive priority queues with multiple servers and two priority classes. *European Journal of Operational Research*, 118(1):181–193.

Karam, M. and Tobagi, F. (2002). Analysis of delay and delay jitter of voice traffic in the internet. *Computer Networks*, 40(6):711–726.

Katsaros, A. and Langaris, C. (1995). An N-class structured priority queue with vacations. *Communications in Statistics - Stochastic Models*, 11(2):235–248.

Khamisy, A. and Sidi, M. (1992). Discrete-time priority queues with two-state markov modulated arrivals. *Stochastic Models*, 8(2):337–357.

Kleinrock, L. (1976). *Queueing systems volume II: Computer applications*. John Wiley & Sons.

Kouvatsos, D. and Awan, I. (2003). Entropy maximisation and open queueing networks with priorities and blocking. *Performance Evaluation*, 51(2-4):191–227.

Kouvatsos, D. and Tabet-Aouel, N. (1994). An ME-based approximation for multi-server queues with preemptive priority. *European Journal of Operational Research*, 77(3):496–515.

Kraimeche, B. (2001). Multiplexing of video and data sources in an ATM access network. *Computer Communications*, 24(9):889–897.

Krinik, A., Marcus, D., Shiflett, R., and Chu, L. (2002). Transient probabilities of a single server priority queueing system. *Journal of Statictical Planning and Inference*, 101(1-2):185–190.

Laevens, K. (1999). *Stochastische modellering van ATM-schakelelementen met buffers aan de ingangszijde*. PhD thesis, Ghent University.

Laevens, K. and Bruneel, H. (1998). Discrete-time multiserver queues with priorities. *Performance Evaluation*, 33(4):249–275.

Langaris, C. and Katsaros, A. (1995). Time-dependent analysis of a queue with batch arrivals and n levels of nonpreemptive priority. *Queueing Systems*, 19(3):269–288.

Langaris, C. and Katsaros, A. (1997). Preemptive resume priorities in an N-class structured queue with server vacations. *Journal of the Operations Research Society of Japan*, 40(1):45–61.

Lee, J., Choi, J., and Un, C. (1998). Performance analysis of an input queueing ATM switch with two priority classes. *Performance Evaluation*, 32(2):137–149.

Lee, Y. (2001). Discrete-time $\text{Geo}^{\text{X}}/\text{G}/1$ queue with preemptive resume priority. *Mathematical and Computer Modelling*, 34(3-4):243–250.

Lee, Y. and Choi, B. (2001). Queueing system with multiple delay and loss priorities for ATM networks. *Information Science*, 138(1-4):7–29.

Lee, Y., Kim, Y., and Huh, J. (2003). Discrete-time $\text{Geo}^{\text{X}}/\text{G}/1$ queue with non-preemptive priority. *Computers and Mathematics with Applications*, 46(10-11):1625–1632.

Leemans, H. (2001). Waiting time distribution in a two-class two-server heterogeneous priority queue. *Performance Evaluation*, 43(2-3):133–150.

Liebeherr, J. and Wrege, D. (1999). Priority queue schedulers with approximate sorting in output-buffered switches. *IEEE Journal on Selected Areas in Communications*, 17(6):1127–1144.

Little, J. (1961). A proof for the queueing formula $l = \lambda w$. *Operations Research*, 9:383–387.

Liu, K., Petr, D., Frost, V., Zhu, H., Braun, C., and Edwards, W. (1997). Design and analysis of a bandwidth management framework for atm-based broadband ISDN. *IEEE Communications Magazine*, pages 138–145.

Liu, Y. and Gong, W. (2003). On fluid queueing systems with strict priority. *IEEE Transactions on Automatic Control*, 48(12):2079–2088.

Machihara, F. (1995). A bridge between preemptive and nonpreemptive queueing models. *Performance Evaluation*, 23(2):93–106.

Maertens, T., Walraevens, J., and Bruneel, H. (2004). Performance analysis of a single-server queue with HOL-PJ priority scheduling discipline. In *Proceedings of the Second International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs '04*, pages P42/1–P42/10, Ilkley.

Mandjes, M. (2003). Pricing strategies under heterogeneous service requirements. *Computer Networks*, 42(2):231–249.

Marks, B. (1973). State probabilities of M/M/1 priority queues. *Operations Research*, 21(4):974–987.

Mehmet Ali, M. and Song, X. (2004). A performance analysis of a discrete-time priority queueing system with correlated arrivals. *Performance Evaluation*, 57(3):307–339.

Meir, A. and Moon, J. (1989). On an asymptotic method in enumeration. *Journal of Combinatorial Theory, series A*, 51:77–89.

Miller, D. (1981). Computation of steady-state probabilities for M/M/1 priority queues. *Operations Research*, 29(5):945–958.

Miller, R. (1960). Priority queues. *Annals of Mathematical Statistics*, 31:86–103.

Mitrani, I. and King, P. (1981). Multiprocessor systems with preemptive priorities. *Performance Evaluation*, 1(2):118–125.

Mukherjee, S., Saha, D., and Tripathi, S. (1995). A preemptive protocol for voice-data intergration in ring-based lan: performance analysis and comparison. *Performance Evaluation*, 11(3):339–354.

Núñez Queija, R. and Boxma, O. (1998). Analysis of a multi-server queueing model of ABR. *Journal of Applied Mathematics and Stochastic Analysis*, 11(3):339–354.

Ozawa, T. (1992). Analysis of a multiqueue model for an ISDN access interface. *Performance Evaluation*, 15(2):65–176.

Parekh, A. and Gallager, R. (1994). A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150.

Paterok, M. and Ettl, M. (1994). Sojourn time and waiting time distributions for m/gi/1 queues with preemption-distance priorities. *Operations Research*, 42(6):1146–1161.

Peterson, W. (1991). A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research*, 16(1):90–118.

Rubin, I. and Tsai, Z. (1989). Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems. *IEEE Transactions on Information Theory*, 35(3):637–647.

Sandhu, D. and Posner, M. (1989). A priority M/G/1 queue with application to voice/data communication. *European Journal of Operational Research*, 40(1):99–108.

Schaack, C. and Larson, R. (1986). An N-server cutoff priority queue. *Operations Research*, 34(2):257–266.

Schormans, J., Pitts, J., and Scharf, E. (1991). Time priorities in ATM switches. In *Proceedings of ITC 13*, pages 527–532, Copenhagen.

Shakkottai, S. and Srikant, R. (2001). Many-sources delay asymptotics with applications to priority queues. *Queueing Systems*, 39(2-3):183–2000.

Sharma, V. and Virtamo, J. (2002). A finite buffer queue with priorities. *Performance Evaluation*, 47(1):1–22.

Shenker, S. (1995). Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications*, 13(7):1176–1188.

Sidi, M. (1987). Discrete-time priority queues with partial interference. *IEEE Journal on Selected Areas in Communications*, 5(6):1041–1050.

Sidi, M. (1988). Two competing discrete-time queues with priority. *Queueing Systems*, 3(4):347–362.

Sidi, M. and Segall, A. (1983). Structured priority queueing systems with applications to packet-radio networks. *Performance Evaluation*, 3(4):265–275.

Stanford, D. (1991). Interdeparture-time distributions in the non-preemptive priority $\sum M_i/G_i/1$ queue. *Performance Evaluation*, 21(1):43–60.

Stanford, D. (1997). Waiting and interdeparture times in priority queues with poisson- and general-arrival streams. *Operations Research*, 45(5):725–735.

Subramanian, V. and Srikant, R. (2000). Tail probabilities of low-priority waiting times and queue lengths in MAP/GI/1 queues. *Queueing Systems*, 34(1-4):215–236.

Sugahara, A., Takine, T., Takahashi, Y., and Hasegawa, T. (1995). Analysis of a nonpreemptive priority queue with SPP arrivals of high class. *Performance Evaluation*, 21(3):215–238.

Sumita, U. and Sheng, O. (1988). Analysis of query processing in distributed database systems with fully replicated files: a hierarchical approach. *Performance Evaluation*, 8(3):223–238.

Tabet-Aouel, N. and Kouvatsos, D. (1992). On an approximation to the mean response times of priority classes in a stable G/G/c/PR queue. *Journal of the Operational Research Society*, 43(3):227–239.

Takada, H. and Miyazawa, M. (2002). A Markov modulated fluid queue with batch arrivals and preemptions. *Stochastic Models*, 18(4):529–652.

Takagi, H. (1990). Priority queues with setup times. *Operations Research*, 38(4):667–677.

Takagi, H. (1991). *Queueing analysis: a foundation of performance evaluation volume 1: vacation and priority systems, part 1*. North-Holland.

Takahashi, M., Ōsawa, H., and Fujisawa, T. (1999). $Geo^{[X]}/G/1$ retrial queue with non-preemptive priority. *Asia-Pacific Journal of Operational Research*, 16(2):215–234.

Takahashi, Y. and Hashida, O. (1991). Delay analysis of discrete-time priority queue with structured inputs. *Queueing Systems*, 8(2):149–164.

Takahashi, Y. and Miyazawa, M. (1994). Relationship between queue-length and waiting time distributions in a priority queue with batch arrivals. *Journal of the Operations Research Society of Japan*, 37(1):48–63.

Takine, T. (1996). A nonpreemptive priority MAP/G/1 queue with two classes of customers. *Journal of Operations Research Society of Japan*, 39(2):266–290.

Takine, T. (1999). The nonpreemptive priority MAP/G/1 queue. *Operations Research*, 47(6):917–927.

Takine, T. and Hasegawa, T. (1994). The workload in the MAP/G/1 queue with state-dependent services: its application to a queue with preemptive resume priority. *Communications in Statistics - Stochastic Models*, 10(1):183–204.

Takine, T., Matsumoto, Y., Suda, T., and Hasegawa, T. (1994a). Mean waiting times in nonpreemptive priority queues with Markovian arrival and i.i.d. service processes. *Performance Evaluation*, 20:131–149.

Takine, T., Sengupta, B., and Hasegawa, T. (1994b). An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. *IEEE Transactions on Communications*, 42(2-4):1837–1845.

Takine, T., Takagi, H., Takahashi, Y., and Hasegawa, T. (1990). Analysis of asymmetric single-buffer polling and priority systems without switchover times. *Performance Evaluation*, 11(4):253–264.

Tham, C., Yao, Q., and Jiang, Y. (2002). Achieving differentiated services through multi-class probabilistic priority scheduling. *Computer Networks*, 40(4):577–593.

Tijms, H. (1974). A control policy for a priority queue with removable server. *Operations Research*, 22(4):833–837.

van der Heijden, M., Van Harten, A., and Sleptchenko, A. (2004). Approximations for Markovian multi-class queues with preemptive priorities. *Operations Research Letters*, 32(3):273–282.

van der Mei, R., van den Berg, J., Vranken, R., and Gijsen, B. (2003). Sojourn-time approximations for a multi-server processor sharing system with priorities. *Performance Evaluation*, 54(3):249–261.

Venkataramani, B., Bose, S., and Srivathsan, K. (1997). Queuing analysis of a non-pre-emptive MMPP/D/1 priority system. *Computer Communications*, 20(11):999–1018.

Vinck, B. and Bruneel, H. (1995). A note on the system contents and cell delay in FIFO ATM-buffers. *Electronics Letters*, 31(12):952–954.

Wagner, D. (1997). Waiting times of a finite-capacity multi-server model with non-peemptive priorities. *European Journal of Operational Research*, 102(1):227–241.

Walraevens, J. and Bruneel, H. (1999). HOL priority in an ATM output queueing switch. In *Proceedings of the seventh IFIP workshop on performance modelling and evaluation of ATM/IP networks*, Antwerp.

Walraevens, J., Steyaert, B., and Bruneel, H. (2000a). Analysis of a GI-Geo-1 preemptive resume priority buffer. In *Proceedings of the IFIP ATM & IP 2000 Workshop*, pages 88/1–88/11, Ilkley.

Walraevens, J., Steyaert, B., and Bruneel, H. (2000b). *Analysis of packet delay in a GI-G-1 queue with non-preemptive priority scheduling*, pages 433–445. LNCS 1815 (Proceedings of the Networking 2000 Conference, Paris). Springer Verlag.

Walraevens, J., Steyaert, B., and Bruneel, H. (2000c). Analysis of the system contents in a GI-G-1 queue with non-preemptive priority scheduling. In *Book of Abstracts ORBEL 14*, pages 32–33, Mons.

Walraevens, J., Steyaert, B., and Bruneel, H. (2000d). Performance analysis of the system contents in a discrete-time non-preemptive priority queue with general service times. *Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL)*, 40(1-2):91–103.

Walraevens, J., Steyaert, B., and Bruneel, H. (2001). Analysis of a preemptive resume priority buffer with general service times for the high priority class. In *Proceedings of the Africom 2001 Conference*, Cape Town.

Walraevens, J., Steyaert, B., and Bruneel, H. (2002a). Delay characteristics in discrete-time GI-G-1 queues with non-preemptive priority queueing discipline. *Performance Evaluation*, 50(1):53–75.

Walraevens, J., Steyaert, B., and Bruneel, H. (2002b). *Performance analysis of a GI-G-1 preemptive resume priority buffer*, pages 745–756. LNCS 2345 (Proceedings of the Networking 2002 Conference, Pisa). Springer Verlag.

Walraevens, J., Steyaert, B., and Bruneel, H. (2003a). Analysis of a preemptive repeat priority buffer with resampling. In *Proceedings of the International Network Optimization Conference (INOC 2003)*, pages 581–586, Evry.

Walraevens, J., Steyaert, B., and Bruneel, H. (2003b). *Delay analysis of a discrete-time non-preemptive priority buffer with 3 traffic classes*, pages 350–357. Recent Advances in Communications and Computer Science. WSEAS Press.

Walraevens, J., Steyaert, B., and Bruneel, H. (2003c). Performance analysis of a single-server ATM queue with a priority scheduling. *Computers & Operations Research*, 30(12):1807–1829.

Walraevens, J., Steyaert, B., and Bruneel, H. (2004a). A packet switch with a priority scheduling discipline: Performance analysis. *Telecommunication Systems*. To appear.

Walraevens, J., Steyaert, B., and Bruneel, H. (2004b). Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline. *European Journal of Operational Research*, 157(1):130–151.

Walraevens, J., Steyaert, B., and Bruneel, H. (2004c). Performance evaluation of a discrete-time HOL priority queue with multiple traffic classes. Technical report, Ghent University.

Wang, J., Hwang, W., Wang, W., and Shieh, C. (2000). Performance evaluation of multiple-channel slotted ring networks with tunable transmitters and fixed receivers. *Computer Communications*, 23(13):1281–1291.

Wang, Q. (2004). Modeling and analysis of high risk patient queues. *European Journal of Operational Research*, 155(2):502–515.

White, H. and Christie, L. (1958). Queuing with preemptive priorities or with breakdown. *Operations Research*, 6(1):79–95.

Williams, T. (1980). Nonpreemptive multi-server priority queues. *The Journal of the Operational Research Society*, 31(12):1105–1107.

Wittevrongel, S. (1998). *Prestatie-analyse van discrete-tijd-buffersystemen met grillige aankomststromen*. PhD thesis, Ghent University.

Xabier Albizuri, F., Graña, M., and Raducanu, B. (2003). Statistical transmission delay guarantee for nonreal-time traffic multiplexed with real-time traffic. *Computer Communications*, 26(12):1365–1375.

Xiao, X. and Ni, L. (1999). Internet QoS: a big picture. *IEEE Network*, 13(2):8–18.

Xiong, Y. and Bruneel, H. (1993). Buffer contents and delay for statistical multiplexers with fixed-length packet-train arrivals. *Performance Evaluation*, 17(1):31–42.

Yoon, C. and Un, C. (1991). Unslotted CSMA-CD protocols with combined retransmission strategy for fiber optic bus and ring networks. *Computer Networks and ISDN Systems*, 21(5):381–397.

Yoon, C. and Un, C. (1994). Unslotted 1- and p_i-persistent CSMA-CD protocols for fiber optic bus networks. *IEEE Transactions on Communications*, 42(2-4):158–465.