Processor-Sharing Models for Integrated-Services Networks

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof.dr. M. Rem,
voor een commissie aangewezen door het College
voor Promoties in het openbaar te verdedigen
op donderdag 20 januari 2000 om 16.00 uur

door

Rudesindo Núñez Queija

geboren te Heemskerk

Contents

Summary

In this thesis we study *queueing* models which can be used in the performance analysis of *integrated-services* telecommunication networks. Chapter 1 gives an overview of the evolution of these networks and describes the most relevant features. Modern telecommunication systems offer a wide range of services (data, voice, video) which are carried simultaneously in the network on an integrated basis. We can roughly divide the traffic into two broad classes: *stream* traffic and *elastic* traffic. Stream traffic mainly consists of "real-time" connections (such as telephony and interactive video applications) which are extremely sensitive to transmission delays. Stream connections therefore require a certain guaranteed capacity. Elastic traffic (data transmission, e-mail) on the other hand allows for fluctuations in the transmission rate, as long as the total delay is "acceptable". The transmission capacity available to elastic traffic varies as stream traffic connections are set up or terminated. Each elastic traffic connection gets an equal share of the capacity left over by stream traffic. In the thesis we focus on the performance analysis of elastic traffic, using so-called processor-sharing models with varying service capacity. An elastic traffic connection is represented by a customer in a queueing model. Hence, the service requirement of a customer in the model corresponds to the size of, for instance, a data file. The service capacity, which fluctuates according to some stochastic process, is shared among the customers in the queue according to the processor-sharing discipline, i.e., each customer gets an equal share. Processor-sharing models with *constant* service capacity are well-studied in the literature. Fluctuations in the service capacity, however, turn out to make the analysis considerably more complicated. This thesis presents the first analytic results concerning *sojourn* times in processor-sharing queues with varying service capacity (which correspond to the transmission times of elastic services).

In Chapter 2 we first derive the queue-length distribution in an M/M/1 (processor-sharing) queue of which the service capacity (and arrival intensity) varies depending on the state of a birth-death process. The queue-length distribution is obtained by combining the theory of matrix-geometric solutions with the method of spectral expansion. The theory of matrix-geometric solutions enables a transparent analysis using probabilistic arguments, while the spectral expansion allows for a more detailed analysis. We also show how the alternative method of generating functions can be applied, and we discuss the intimate relation between the three approaches. Special attention is devoted to the in-

fluence of the (capacity) fluctuations when these occur either very fast or very slow (relative to the service times of customers). We show that approximating the system by one with constant service capacity, equal to the average service capacity in the model with fluctuations, is only justified when the fluctuations occur very fast (so that they average out). The formal analysis is illustrated by numerical experiments for a specific telecommunication system.

In the remainder of the thesis we concentrate on the sojourn times of customers, in particular conditional on the service requirement. In Chapter 3 we study a processor-sharing model of which the service capacity is constant during so-called *on-periods* and no service is rendered during *off-periods*. We again assume that the service requirements have an exponential distribution. The sojourn time distribution is given in terms of its LST (Laplace-Stieltjes Transform). The analysis is based on a random time-scale transformation, via which sojourn times in the original model are represented by transient rewards in a branching process with a specific reward structure. We further show that the decomposition of the sojourn time into *independent* components, which is known for processor-sharing models with constant service capacity, also applies to the on/off model. Another well-known property of standard processor-sharing models is that the expected conditional sojourn time is a linear function of the service requirement. In the on/off model it turns out that this is only true asymptotically, that is, for large service requirements.

In Chapter 4 we study sojourn times in the case that the service capacity depends on the state of a general Markov process. In contrast to the on/off model, service can be rendered at *different* positive rates. This generalisation prohibits an analysis as detailed as the one presented for the on/off model. In particular, the above mentioned decomposition of sojourn times no longer applies. However, the asymptotic linearity of the expected conditional sojourn time as a function of the service requirement is preserved. This is shown using the LST of the conditional sojourn time, which is again derived using the method of time-scale transformation. We also discuss *why* the above mentioned linearity is lost when the service capacity fluctuates. The results of the analysis are then used in numerical experiments for the performance evaluation of a communication system. The analytic and numerical results lead to a good and simple approximation of the expected conditional sojourn time. The analysis can be extended to the case that the service requirements have a phase-type distribution. Furthermore, the analysis also applies to the more general service discipline *discriminatory* processor sharing. Both generalisations, however, are at the expense of a higher computational complexity.

In Chapter 5 we study the tail of the sojourn time distribution in the case that the service requirement distribution has a so-called *heavy tail*. It is well-known that when the latter is the case and customers are served in the order of arrival (the so-called *First Come First Served* discipline), then the tail of the sojourn time distribution is "one degree" heavier: it is as heavy as the *integrated* tail of the service requirement distribution. As a consequence, the mean sojourn time is infinite when the variance of the service requirements is infinite. It is also known that with the processor-sharing discipline (and constant service

capacity) the tails of the sojourn time and the service requirement distributions are exactly as heavy. This is generally seen as a desirable property. We generalise this result to the on/off model assuming a heavy-tailed service requirement distribution (this was not the case in Chapter 3). We do so by generalising the decomposition property of the sojourn times in the on/off model to the case of generally distributed service requirements. The approach also leads to a new and simpler proof of the result in the standard processor-sharing model. Furthermore, we establish the "tail equivalence" of the sojourn time and service requirement distributions for two other disciplines: *foreground-background processor sharing* (only the customers that have received the least amount of service are served in processor-sharing fashion), and *shortest remaining processing time first* (in which the customers with the smallest remaining service requirement are served).

Samenvatting

In dit proefschrift worden *wachtrijmodellen* bestudeerd, die gebruikt kunnen worden in de prestatie-analyse van telecommunicatiesystemen met *geïntegreerde* diensten. Hoofdstuk 1 geeft achtergrondinformatie over de historische ontwikkeling van telecommunicatiesystemen met geïntegreerde diensten en beschrijft de meest relevante eigenschappen. Moderne communicatiesystemen bieden de mogelijkheid om simultaan zeer verschillende typen verkeer (data, geluid, video) in geïntegreerde vorm over hetzelfde netwerk te versturen. We kunnen de verschillende soorten verkeer ruwweg indelen in twee klassen: *stroom* verkeer en *elastisch* verkeer. Stroom verkeer bestaat voornamelijk uit "real-time" verbindingen (o.a. telefonie en interactieve video-applicaties) die nauwelijks vertragingen in de transmissie tolereren; derhalve is voor die diensten een zekere capaciteitsgarantie vereist. Elastisch verkeer (o.a. datatransmissie, e-mail) daarentegen laat fluctuaties in de transmissiesnelheid toe, zolang de *totale transmissieduur* "acceptabel" is. Als gevolg van het opzetten en afbreken van connecties van stroom verkeer, varieert de transmissiecapaciteit die beschikbaar is voor elastisch verkeer. Elke connectie van elastisch verkeer deelt in gelijke mate in de capaciteit die overgelaten wordt door het stroom verkeer. In het proefschrift gaat de aandacht uit naar de prestatie-analyse van elastisch verkeer door middel van zogenaamde *processor-sharing* modellen met variërende capaciteit. Een connectie van een elastische dienst wordt gerepresenteerd door een klant in een wachtrijmodel. De bedieningsvraag van de klant in het wachtrijmodel correspondeert dus bijvoorbeeld met de omvang van een data bestand in het oorspronkelijke communicatiesysteem. De bedieningscapaciteit, die fluctueert volgens een stochastisch proces, wordt op elk moment gelijk verdeeld ("processor sharing") onder de aanwezige klanten. Processor-sharing modellen met *constante* bedieningscapaciteit zijn reeds uitvoerig bestudeerd. De fluctuerende capaciteit blijkt echter een belangrijke complicerende factor te zijn in de analyse. In dit proefschrift worden voor het eerst analytische resultaten verkregen voor de verdeling van *verblijftijden* van klanten in zo'n systeem.

In Hoofdstuk 2 wordt allereerst de verdeling van het aantal klanten in een M/M/1 (processor-sharing) wachtrij bepaald, waarbij de bedieningscapaciteit (*en* de aankomstintensiteit) varieert volgens een geboorte-sterfte proces. De rijlengte verdeling wordt verkregen door middel van resultaten van de theorie van matrix-geometrische oplossingen in combinatie met de techniek van spectrale ontwikkeling. De theorie van matrix-geometrische oplossingen maakt de

afleiding inzichtelijk door het gebruik van probabilistische argumenten, terwijl de spectrale ontwikkeling een gedetailleerdere analyse mogelijk maakt. Tevens wordt aangegeven hoe, als alternatief, de methode van genererende functies gebruikt kan worden en wordt de relatie tussen de drie verschillende methoden besproken. Speciale aandacht wordt geschonken aan het effect van de fluctuaties wanneer deze zeer snel of juist zeer traag plaats vinden (ten opzichte van de verblijftijd van klanten). Aangetoond wordt dat de benadering door middel van een systeem met *constante* bedieningscapaciteit en aankomstintensiteit gelijk aan de overeenkomstige *gemiddelde* waarden in het model met fluctuaties, alleen gerechtvaardigd is wanneer de fluctuaties zeer snel plaats vinden (waardoor uitmiddeling optreedt). De formele analyse wordt geïllustreerd met behulp van numerieke experimenten voor een specifiek telecommunicatiesysteem.

In de rest van het proefschrift concentreren we ons op de verblijftijd van klanten (dit correspondeert met de transmissieduur van elastische diensten), in het bijzonder geconditioneerd op de bedieningsvraag. In Hoofdstuk 3 bestuderen we een processor-sharing model waarbij de bedieningscapaciteit constant is gedurende zogenaamde *aan-periodes* van de bediende en er geen bediening is gedurende *uit-periodes*. We nemen wederom aan dat de bedieningsvraag exponentieel verdeeld is. De kansverdeling van de conditionele verblijftijd wordt gegeven in termen van de LST (Laplace-Stieltjes Transformatie). Hiervoor wordt, door middel van een (stochastische) tijdschaal-transformatie, het verblijftijden-probleem geformuleerd in termen van een vertakkingsproces met een specifieke opbrengsten-structuur. We tonen verder aan dat de — voor processor-sharing modellen met constante capaciteit — bekende decompositie van de verblijftijd in *onafhankelijke* componenten, behouden blijft in het aan/uit-model. Een andere eigenschap van standaard processor-sharing modellen (met constante capaciteit) is dat de verwachte conditionele verblijftijd een lineaire functie is van de bedieningsvraag. Voor het aan/uit-model blijkt deze eigenschap echter alleen asymptotisch (voor grote bedieningsvraag) te gelden.

In Hoofdstuk 4 bestuderen we de verblijftijden in een model waarbij de bedieningscapaciteit afhangt van de toestand van een algemeen Markov proces. Anders dan in het aan/uit-model kan de bedieningscapaciteit *verschillende* positieve waarden aannemen. Deze generalisatie staat een gedetailleerde analyse zoals in het aan/uit-model in de weg. In het bijzonder blijkt de bovengenoemde decompositie van de verblijftijd in onafhankelijke componenten niet langer te gelden. De asymptotische lineariteit van de verwachte conditionele verblijftijd blijft echter wel behouden. Dit wordt aangetoond met behulp van de LST van de conditionele verblijftijd-verdeling, die wederom gevonden wordt door middel van een tijdschaal-transformatie. Ook wordt verklaard *waarom* de lineariteit verstoord wordt, wanneer de bedieningscapaciteit fluctueert. Door middel van numerieke experimenten worden de verkregen resultaten toegepast in de prestatie-analyse van een specifiek communicatiesysteem met stroom en elastisch verkeer. De analytische en numerieke resultaten leiden tot een goede en eenvoudige benadering van de verwachte conditionele verblijftijd. De analyse kan gegeneraliseerd worden naar het geval dat de bedieningsvraag een *fase-type* verdeling heeft. Ook geldt de analyse voor hetzelfde model met de

algemenere bedieningsdiscipline *discriminatory* processor-sharing. Beide generalisaties brengen echter een hogere numerieke complexiteit met zich mee.

In Hoofdstuk 5 bestuderen we de staart van de verblijftijd-verdeling in het geval dat de bedieningsvraag-verdeling een zogenaamde *zware staart* heeft. Het is bekend dat als dit laatste het geval is en klanten in volgorde van aankomst bediend worden (de zogenaamde *First Come First Served* discipline), dan is de staart van de verblijftijd-verdeling "één graad" zwaarder dan die van de bedieningsvraag-verdeling (namelijk even zwaar als de geïntegreerde staart van de laatst genoemde). Hierdoor is bijvoorbeeld de verwachte verblijftijd oneindig wanneer de variantie van de bedieningsvraag oneindig is. Ook is bekend dat onder de processor-sharing discipline geldt dat de staarten *precies even zwaar* zijn, wat in het algemeen gezien wordt als een wenselijke eigenschap. Dit laatste resultaat generaliseren we voor het aan/uit-model waarbij de bedieningsvraagverdeling een zware staart heeft (in Hoofdstuk 3 was dat niet het geval). Hiervoor generaliseren we onder meer de decompositie eigenschap van de verblijftijd voor het geval dat de bedieningsvraag in het aan/uit-model een algemene verdeling heeft. De gekozen aanpak leidt tevens tot een eenvoudiger bewijs van het reeds bekende resultaat in het gewone processor-sharing model (met constante capaciteit). Met behulp van dezelfde bewijstechniek wordt de eigenschap ook bewezen voor twee andere bedieningsdisciplines: *foreground-background processor sharing* (waarbij de klanten met de minste reeds verkregen bediening volgens processor sharing worden bediend) en *shortest remaining processing time first* (waarbij de klanten met de minste resterende hoeveelheid werk eerst worden bediend).

About the author/Over de auteur

Sindo (Rudesindo) Núñez Queija was born in Heemskerk (The Netherlands) on May 10, 1972. He graduated from Grammar School (Augustinus College, Beverwijk) on June 13, 1990. For two years he was a student assistant in Operations Research at the Econometrics Department of the Vrije Universiteit in Amsterdam. He also participated in the European project *Human Capital and Mobility* at the Universitat Politècnica de Catalunya (Barcelona, Spain). He received his master's degree in econometrics (cum laude) from the Vrije Universiteit (Amsterdam) on June 29, 1995, after which he became research assistant at CWI (Centre for Mathematics and Computer Science, Amsterdam). Since August 1, 1999, he is a post-doc at CWI. He defends this PhD thesis at the Technische Universiteit Eindhoven on January 20, 2000.

Sindo (voluit: Rudesindo) Núñez Queija werd geboren op 10 mei 1972 in Heemskerk. Op 13 juni 1990 behaalde hij het VWO diploma aan het Augustinus College te Beverwijk, waarna hij econometrie ging studeren aan de Vrije Universiteit van Amsterdam. Daar was hij twee jaar lang student-assistent in de Operations Research (vakgroep Econometrie). Verder werkte hij een half jaar in het Europese project *Human Capital and Mobility* aan de Universitat Politècnica de Catalunya (Barcelona, Spanje). Op 29 juni 1995 behaalde hij het doctoraal diploma Econometrie (cum laude). Hierna werd hij onderzoeker-in-opleiding aan het Centrum voor Wiskunde en Informatica in Amsterdam. Sinds 1 augustus 1999 is hij daar werkzaam als post-doc. Op 20 januari 2000 verdedigt hij dit proefschrift aan de Technische Universiteit Eindhoven.