# Processor-sharing models for GPRS communication networks

van Pelt, M.

*Award date:*
2009

TECHNISCHE UNIVERSITEIT EINDHOVEN

Department of Mathematics and Computer Science

MASTER'S THESIS

# Processor-Sharing Models for GPRS Communication Networks

by
M.v.Pelt

Supervisors: prof.dr.ir. S.C. Borst and dr. R. Núñez Queija

Eindhoven, August 17, 2009

# Contents

# Acknowledgements

This thesis would never have been completed without the support of various people. For their support I would like to express my gratitude. At first I want to thank Rudesindo (Sindo) Núñez Queija and Sem Borst. They both were my supervisors from Eindhoven University of Technology (TU/e). In the first part of this project Sindo helped me with his knowledge and personal guidance. His quick and accurate feedback on my work helped me gain insight in the Processor Sharing queue. In the second part of the project Sem Borst became my supervisor. He helped me to regain motivation to finish the project. His corrections and comments always came with such speed that I was able to make progress quickly. The conversations with Sem were pleasant and helped me to solve the mean conditional sojourn time in the admission controlled Processor Sharing queue. He gave me a different view on the model by pointing out the properties of symmetric queues. Also I would like to thank both Sem and Sindo for supporting me as a person and giving me the chance to finish my Mathematics study at the TU/e. Another person I would like to thank is Jeroen Wieland from Vodafone. He was my supervisor during my stay at Vodafone in Maastricht. Without him this project would never have started at all. He gave me insight in the world of GPRS-systems and mobile phones. He was very interested in the mathematics of this project, but always reminded me that the project wasn't purely about mathematics. The dimensioning of the system, and so one of the major goals of this thesis, were part of the project he supervised as researcher at Vodafone Maastricht. At the TU/e Onno Boxma helped me with a literature study that I did on the side of this project. He helped me by his kindness and gave me the feeling this project could be a success. From him I also got the motivation to finally finish this project. There is also a word of gratitude for Cor Hurkens and Matthieu Jonckheere who agreed to serve as members of in the graduation committee of this project. Altogether, without the kindness of all the people at the mathematics department of the TU/e I would have never got the confidence and motivation needed. The last person I would like to thank is Heidi van Beurden. She is the one person who always supported me. She never let me down, even in hard times. Without her this project would have ended down the drain. Heidi showed me that I could do things I never imagined. All of you persons are in my heart!

Merijn van Pelt, July 2009.

# Chapter 1

# Introduction

In the last two decades wireless telecommunication has experienced major growth, especially in the area of mobile telephony. With the deployment of the world's first public wireless telephony network in Tokyo in 1979, the first step towards a global commercial telecommunication system was made. At first the systems used analogue technologies, but with the rapidly growing market and the emerging demand for mobile data services the technology shifted towards digital systems. Of the second-generation mobile communication systems, Global System for Mobile Communications (GSM) is used initially in Europe and parts of Asia. Other parts of the world adopted different 2G systems. Around 1992 the commercial launch of the second-generation GSM took place and in the first quarter of 2004, already accounting for more than 70% of world's wireless market, its billionth user was connected. With the rise of the Internet, more and more data services appeared, demanding ever higher bitrates and fast access ('always online'). These needs were mostly fulfilled by the introduction of General Packet Radio Service (GPRS), which enabled the dynamic and flexible sharing of multiple traffic channels by multiple data flows. GPRS shares GSM frequency bands with telephone and circuit-switched data traffic, and makes use of many of the properties of the physical layer of the original GSM system, most importantly the time-division multiple access (TDMA) frame structure, modulation technique, and structure of GSM timeslots. With GPRS also came the possibility of sensible charging on a per volume basis, which makes it affordable to be always online.

The main problem in this thesis is the dimensioning of the GPRS network. We will mathematically model the system in order to derive measures of QoS constraints like throughput and jitter of a data flow in the integrated voice and data telecommunication system. Our focus is to model a single communication cell in the network through the use of Processor-Sharing models.

The outline of this chapter is as follows. In sections 1.1 and 1.2 we will first give a brief historic overview of telecommunication systems. This historic overview is taken from [13]. For a more elaborate overview of the early history of wireless communications, we refer to [1]. In section 1.3 the technologies and standards of interest when modeling the system will be discussed in greater detail and the concept of Quality of Service is introduced. Finally, in section

1.4 we will formulate the goal of this thesis.

## 1.1 Wireless telecommunication systems

In 1864, James Clerk Maxwell postulated the possibility of generating electromagnetic waves that would propagate at the speed of light, which was subsequently demonstrated by Heinrich Rudolph Hertz in 1887. Nikola Tesla was the first to publicly demonstrate wireless transmission in 1893 and shortly thereafter, in 1895, Alexander Stepanovich Popov and Guglielmo Marconi independently demonstrated the electromagnetic transmission and reception of messages. The world's first patent on wireless telegraphy using Hertzian waves was awarded to Marconi in 1896 but overturned by the U.S. Supreme Court in 1943 in favour of Nikola Tesla, after 30 years of legal battles.

The earliest experiments with wireless telephony were done by Reginald Aubrey Fessenden. Where Marconi generated spark-based signals that could be used in Morse code telegraphy, Fessenden recognised that continuous wave transmission was required for speech telephony. On December 23, 1900, he generated the first-ever intelligible speech successfully broadcast by radio waves. The quality was still poor and the distance short, but over the years wireless speech telephony was further enhanced. Initially wireless communication used Amplitude Modulation (AM) and later the more robust Frequency Modulation (FM) scheme (developed by Edwin Howard Armstrong in 1935) was introduced, which formed the technological basis for the first analogue cellular networks.

## 1.2 Evolution of mobile networks

In 1979, the Japanese telecommunications operator NTT deployed the world's first public wireless telephony network in Tokyo. In Europe, mobile telephony was introduced when the Nordic Mobile Telephone (NTM) systems became operational in Scandinavia. Slightly different versions were later on spread through different countries in Europe. Also, in the United Kingdom, there was the Total Access Communications System (TACS) and in the United States the Advanced Mobile Phone System (AMPS) was used. These are typical examples of analogue system used in the 1980s. Collectively, these (analogue) systems are now usually referred to as first-generation (1G) systems which all applied Armstrong's FM-based *analogue* modulation. The systems were mainly used for voice communication, but data communication was also supported. However, due to high error rates and a low bit rate (max. 2.4 kbps without overhead or error correction), very few data applications were practical. Data communication was mostly used for paging and for service engineers to download small data files. Network congestion, areas of poor reception and high degree of compression on voice contributed to low voice quality. Also, most of the first-generation systems were generally unable to interoperate.

In 1982, the Conférence Européenne des administrations des Postes et des Télécommunications (CEPT), installed the Groupe Spécial Mobile, with the task to devise a pan-European mobile telecommunication system. This task

was in 1989 transferred to the European Telecommunications Standards Institute (ETSI). Three years later, the commercial launch of the second-generation (2G) Global System for Mobile communications (GSM) took place. GSM, which is part of the second-generation *digital* systems, offers superior speech quality, international roaming, a high security level, low-power hand-portable terminals and a variety of new services. Still, in 2G there is diversity in the technologies used throughout the world. Europe and parts of Asia adopted the GSM standard, while in the United States and parts of Asia other standards and systems like North American/United States digital communication (NADC/USDC) and Digital Advanced Mobile Phone System (DAMPS), were used. In addition to the technological differences there are also administrative differences between European and the American systems. The subscriber's identity is described differently and they use different protocols for transmitting such information: GSM-Mobility Application Part (GSM-MAP) in Europe, and American National Institute standard 41 (ANSI-41) in the United States. Because of this, a gateway is required to handle administrative matters to let the systems interoperate.

The GSM standard has been further evolved in order to support more services at higher transfer rates. In PHASE 1 the services of speech telephony, Short Message Service (SMS) and Circuit-Switched Data (CSD) can transfer generally at speeds up to 9.6 kbit/s. With the advent of the World Wide Web on the Internet came the need to access it through a mobile phone. This resulted in services as Wireless Application Protocol (WAP) for GSM and I-mode by NTT DoCoMo in Japan.

In order to provide a more satisfactory Internet service, higher data rates and uninterrupted Web access while making a voice call were required. In the PHASE 2+ version of GSM, upgrades of the data transfer capabilities are specified. Among these upgrades was a new channel coding scheme with reduced overhead and hence poorer error protection. This scheme has been standardised to offer up to a 14.4 kbits/s CDS information bit rate. Also there is the possibility to assign multiple traffic channels in parallel to a data call. With this, High-Speed Circuit-Switched Data (HSCSD), higher data transfer rates up to $8 \times 14.4 = 115.2$ kbits/s can be achieved by using up to a maximum of 8 traffic channels in parallel. This data rate is acceptable for a wide range of applications.

Another significant upgrade was made by the introduction of General Packet Radio Service (GPRS) which enables packet-switching in 2G networks. GPRS enables the dynamic and flexible sharing of multiple traffic channels by multiple data flows, in order to enhance service quality and resource efficiency. In a circuit-switched system, when a call is made a certain amount of capacity of the traffic channel is permanently reserved to maintain the call. A circuit-switched call will have a nearly constant bit rate for the duration of the call. So, even when there is no information sent over the traffic channel, its capacity is still used. This can result in inefficient usage of the system's capacity, but can for some applications be desirable. With GPRS also came the possibility of sensible charging on a per volume basis, which makes it affordable to be always online. The maximum (theoretical) data rate in GPRS is $8 \times 21.4 = 171.2$ kbits/s. In

contrast to HSCSD, GPRS requires major soft- and hardware upgrades to the access and core networks.

As a final second-generation network upgrade, Enhanced Data rates for Global Evolution (EDGE) introduces new channel coding and higher-level modulation schemes, designed to boost (HS)CSD and GPRS information data rates up to a technical maximum of $8 \times 59.2 = 473.6$ kbits/s.

## 1.3 Quality of Service

When we want to measure the Quality of Service (QoS) in a communication network it is important to realize that quality is determined by the perception of the user. So, we should identify the global features of quality as perceived by an user and translate them into measurable network constraints. Doing this, we should keep in mind the different kinds of call types. Each call type has its specific demands on the network in order to reach a certain level of quality. Before giving a further elaboration, we will identify the global features that contribute to QoS.

The first feature is access to the network for voice users. When a user requests access to the network there are two main factors that determine the quality: The probability of successful access and the access-time. The probability of successful access should be at least 90%, so 9 out of 10 attempts should be successful. In this thesis we used an even stricter constraint, we want the probability of succesful access to be 0.99. The access-time is less critical and should be no more than a few seconds, [15]. The next essential criterion is that a call should not be interrupted for the intended full duration. Due to handovers and variable signal strength it is much harder to maintain a call in a mobile network as compared to a fixed network. A fast moving user could experience multiple handovers during a single call. The probability of failure during a handover should be very small to guarantee an acceptable QoS, [15]. Finally, during a call there should be no significant glitches, loss of data, or disruption to speech. These are some of the tougher aims to achieve on the network, [15].

For calls involving speech, the next criterion is that the voice quality should be such that it is easily intelligible. In practice this means that the compression algorithm should be adequate, that delay should not vary too much, and that the handover and reception issues above should be reasonable, [15].

For data services, the most important single factor is usually the speed of access, followed by errors, loss of data, misdirection and duplication. The impact the various failures have on the data-service depends on its type, [15].

For GPRS (and Universal Mobile Telecommunications System (UMTS)) the different services have been divided into four main classes. Within a class the services all have similar requirements concerning QoS. The QoS classes are, [15]:

- Conversational;
- Streaming;
- Interactive;

- Background.

We will state the characteristics of each of these classes in the following subsections. For a more detailed description we refer to [15]. In this thesis we will mainly focus on the background class for data traffic and look at the throughput. The throughput of the system could be analysed through the mean conditional sojourn time. How throughput is related to the mean conditional sojourn time is explained in section 2.5. For voice traffic the main QoS constraint considered in this thesis is the blocking probability.

### 1.3.1 Conversational class

This class contains all applications that involve person-to-person communication in real-time, such as videoconferencing and interactive video games. The basic qualities for speech itself are low delay, low jitter (delay-variation), reasonable clarity (codec quality), and absence of echo. In the case of multimedia applications, such as videoconferencing, it is also necessary to maintain correct relative timing to the various media streams. This class is tolerant to some errors, as dropping or corruption of a voice packet lasting for typical 20 ms is unlikely to be detected by a user, [15].

### 1.3.2 Streaming class

The streaming class consists of real-time applications that send information without having a human response. Because of the absence of interaction, there is no longer the need for low delay, but the requirements for jitter and media synchronization remain. Also the error tolerance remains, although some applications can cope with more than others. The removal of the low delay criterion makes it possible to use buffering techniques in the end-user equipment to even out the delay variation, so the acceptable level of network jitter is higher than for the conversational class, [15].

### 1.3.3 Interactive class

This class covers both human and machine communication that request data from another device. The first requirement is that the delay is within the time-out of the application or within reasonable time for human response. This delay does not have to be as low as in the conversational class. The second need is for data integrity, [15].

### 1.3.4 Background class

The background class consists of all applications that either receive data passively or actively request it, but without any immediate need to handle the data. The only requirement is for data integrity, although large file transfers will also require an adequate throughput, [15].

## 1.4 Problem description

Nowadays, with over a billion GSM subscribers and new data services emerging, there is an ever growing need for mobile communication. All over the world new network sites are deployed and/or upgraded to handle more traffic. A proper dimensioning of the network requires finding the needed capacity and parameter settings such that a certain level of QoS is achieved.

When traffic mainly consists of voice calls, as in the early days of telecommunication, dimensioning can be done using the well-known Erlang-B formula, see [8]. The Erlang-B formula gives the probability that a customer does not get a tone (blocking probability), given the number of traffic channels and the offered traffic load. This formula performed well for fixed telecommunication systems with traffic mainly consisting of voice calls. With the subsequent development of mobile telecommunication and the ever growing amount of data traffic, the Erlang-B formula was no longer sufficient.

In a mobile telecommunication network calls are transmitted using a radio link. This link is the physical connection between the fixed network and the mobile phone. The bandwidth is limited by the number of frequencies that can be allocated for communication. Voice traffic is converted into a digital stream that is transmitted at a certain frequency of the radio link. The system could be viewed as a system with a certain number of channels. These channels are also called timeslots.

While in analogue systems an entire channel was allocated to a call for its duration, with digital communication efficient sharing of traffic channels became more and more feasible. In GPRS systems all channels can be used for both voice and data traffic, so dimensioning rules should take both into account. The main difference between data and voice traffic is the allocation of the channels. To meet its specific QoS requirements, each voice call uses one channel for the entire duration of the call. Data traffic is broken up into a stream of packets that traverse the network. These packets are combined with some headers and error correction into a Temporary Block Flow (TBF). The fact that more than one channel can be used and channels can be shared among TBFs, as well as the flexibility of data traffic makes data traffic different from voice traffic and so the Erlang-B formula no longer applies. For a system with only data, a natural substitute is provided by the Processor-Sharing model.

In this thesis we will focus on the problem of finding tools for dimensioning an integrated-services GPRS telecommunication cell. For voice traffic the main QoS constraint is the blocking probability, while for data traffic we want the throughput of an individual user to be above a minimal level. Voice and data traffic share the same capacity, so there is an interaction between the two classes. The capacity is first used to serve voice customers. The left over capacity is given to data customers. One could say that the voice traffic induces randomly varying service capacity for data traffic. We will design a mathematical model which can be used to estimate the QoS constraints and captures the major characteristics of the system. We will use the Quasi-Stationary regime (see [6])

in the model to simplify the effects of the randomly varying service capacity.

# Chapter 2

# Models

In this chapter we will look at ways of modeling circuit-switched and packet-switched communication systems. We will use queueing theory to model both systems in isolation. At first we will only model admission control for circuit-switched systems. In the next chapter we will extend the models to handle admission control for packet-switched systems.

This chapter contains the following parts. In section 2.1 the concepts of circuit-switched and packet-switched communication are discussed. We will look at the different layers of a communication network and note the use of the TCP-protocol to avoid congestion in the network. Section 2.2 is about systems using circuit-switched communication and gives the classical Erlang-loss model for these systems. Next, we will look at the Processor-Sharing model for data transmission in section 2.3. The Processor-Sharing model is explained and we give various reasons to apply this model for modeling packet-switched data traffic. In section 2.4 the $M/M/1-PS$ queueing model without admission control is introduced. Results will be presented for the mean conditional sojourn time and the conditional variance of the sojourn time. In section 2.5 we will look at the throughput of data customers. In this thesis throughput is the QoS constraint when dimensioning the system. We want to make statements about the allowed arrival rate of data customers under the constraint that the throughput of each data customer is at least a certain target value. To model the system we could also combine voice and data traffic in a 2-dimensional Markov model. In section 2.6 we will argue why this model is not used in this thesis. Instead we will show the benefits of the so called quasi-stationary regime and use this to analyse the system later on.

## 2.1 Circuit-switched and packet-switched networks

In this section we will discuss the characteristics of packet-switched and circuit-switched communications. The system we want to model in this thesis consists of both types.

When modeling the traffic in the network it is convenient to recognize a three-level hierarchy. At the highest level, the call level or connection level, connections are being established for the duration of the call. While connected,

the message is fragmented into so-called *packets*, which are transmitted through the network. Networks using this technology are called *packet-switched* networks. These packets contain extra information, so they can be reassembled at a specified destination to recover the original message. In Internet Protocol (IP) networks there is no real notion of connections. The packets flow through the network as independent entities. Still we use the term connection to indicate the information flow from source to destination. The level at which the traffic is observed as individual packets in the network is called the *packet level*. Packets belonging to the same connection are usually not generated as a constant flow, but rather in *bursts*. This gives rise to the burst-level, an intermediate level between the connection level and the packet level.

In a *circuit-switched* environment incoming calls request a fixed amount of bandwidth for a certain amount of time. The amount of bandwidth per call is typically the same for all calls. The continuous bandwidth-spectrum can be transformed through use of technology into a discrete set of communication channels, called timeslots. Each incoming voice call needs to be allocated a timeslot in order to communicate. Since the available bandwidth is limited, the total number of timeslots is limited. Also other technological aspects can limit the number of available timeslots. Hence, there is a limit to the number of calls that can be served at the same time. All calls requesting a channel when there are no free channels are 'blocked'.

To protect the network from congestion, files can be sent over the network using the TCP-protocol. This protocol adapts its sending rate dynamically to avoid possible congestion in the network. The protocol offers packets to the network in bursts. The sending rate is adapted by altering the number of packets in a burst. If no congestion is detected, the number of packets in a burst is increased linearly until the protocol detects congestion. If so, the number of packets in a burst is decreased by a given factor. The bursts generated by the TCP-protocol are so-called TCP-windows. If we look at the traffic flow on the burst level, the traffic consists of TCP-windows of packets.

For data communication the *packet-switched* way of communication is used. For voice calls the QoS settings require *circuit-switched* communication. Although the voice calls are still transformed into digital packets, they are allocated a timeslot for the entire duration of the call. So one could say that it is again circuit-switched communication. To get a better understanding, we will first discuss the system as if it was totally circuit-switched.

## 2.2 Erlang-loss model for circuit-switched calls

Assume that there are $c_v$ timeslots in total available for circuit-switched communication. Also, voice calls arrive according to a Poisson process with parameter $\lambda_v$. At first, we will assume that the holding time of a voice call is exponentially distributed with mean $\frac{1}{\mu_v}$. The system described above is known as the Erlang-loss model. It is a loss model, because of blocking due to the limited number of calls that can be in the system simultaneously. All calls arriving when the system has no free channels are blocked and there are no retrials. This way the

system will be stable even with a load $\rho_v = \frac{\lambda_v}{\mu_v} \geq C_v$. In Kendall's notation, see for instance [5, 11, 12], the system can be denoted as $M/M/c_v/c_v$. Here $c_v$ is the number of 'servers', which equals the number of timeslots in the system.

In equilibrium the flow into state $i-1$ from state $i$ is the same as the flow from state $i-1$ into state $i$. From this we obtain

$$p_v(i-1)\lambda_v = p_v(i)i\mu_v \qquad\qquad i = 1, 2, \cdots, c_v.$$

Hence, together with $\sum_{i=0}^{\infty} p_v(i) = 1$ it is readily verified that

$$p_v(i) = \frac{(\lambda_v/\mu_v)^i/i!}{\sum_{k=0}^{c_v}(\lambda_v/\mu_v)^k/k!} = \frac{\rho_v^i/i!}{\sum_{k=0}^{c_v}\rho_v^k/k!} \qquad i = 0, 1, \cdots, c_v, \tag{2.1}$$

where $\rho_v = \lambda_v/\mu_v$.

Now the *blocking probability* $B(c_v, \rho_v)$ is given by

$$B(c_v, \rho_v) = p_v(c_v) = \frac{\rho_v^{c_v}/c_v!}{\sum_{k=0}^{c_v}\rho_v^k/k!}. \tag{2.2}$$

This formula is also known as the *Erlang-B loss formula*, see [8]. It can be evaluated numerically using the following stable recursion

$$B(c_v, \rho_v) = \frac{\rho_v B(c_v - 1, \rho_v)}{c_v + \rho_v B(c_v - 1, \rho_v)}, \tag{2.3}$$

starting with $B(0, \rho_v) = 1$.

One remarkable result is that the steady-state probabilities do not depend on the holding time distribution other than through its mean. The steady-state probabilities are *insensitive* to the distribution of the holding time. So equation (2.1), with $\rho_v = \lambda_v E[B]$, where $E[B]$ is the mean holding time, also represents the steady-state probabilities of an $M/G/c_v/c_v$ loss system, see [8, 19].

## 2.3 Packet-switched data transmission and Processor-Sharing

We will now discuss a model for a system in which all communication is done in a packet-switched fashion. When a file is served, it occupies the processor and in doing so it uses a part of its capacity. When a server gets multiple service requests, these are handled in a time-sharing fashion. This may be modeled by a *Round Robin* (RR) discipline, see [12, p. 166]. In the RR-discipline the files are served in turn, each for a small time slice. An idealization of the RR-discipline is the *Processor-Sharing* (PS) discipline. In this discipline the capacity of the server is equally shared between all the customers in the system, all customers are served simultaneously and service starts immediately upon arrival. Processor-Sharing could be considered as the limit of the Round Robin discipline, where the service duration (time slice) for each customer approaches zero.

On the connection level the Processor-Sharing discipline is a reasonable approximation for packet-switched communication. Here all files are simultaneously served by one processor and service starts immediately upon arrival. Each file roughly gets the same amount of capacity allocated, which is approximately equal to the total capacity divided by the number of files present at the server.

One could put more detail into the model when looking at the packet level. Scheduling algorithms and other technicalities could be adopted in the model, but in doing so the model becomes intractable. This level of detail is not desirable when using queueing models. Also, on the packet level the Processor-Sharing discipline does not model the system well. Another difficulty, when modeling on the packet level, is capturing the characteristics of the arrival process.

To capture the characteristics of the TCP-protocol in a queueing model, we also use the Processor-Sharing discipline. The reasoning behind this modeling assumption is the following. We assume that the TCP-protocol balances the load generated by the various connections and the processor fairly shares its capacity over the various windows, so each connection roughly gets the same amount of processing time. With this we assume that the window size is nearly equal for each connection, so every TCP connection delivers on average roughly the same amount of work per unit of time. If all TCP connections are set up to transmit large files, this assumption is reasonable. When small files are transmitted, some connections, due to the slow start phase of TCP, will not reach this equilibrium of window sizes. Also, we assume all windows generated by the TCP-protocol from the various connections to be processed at the same time, which leads to modeling this system by the Processor-Sharing discipline.

## 2.4 M/M/1-PS

For the remainder of this thesis we will use the Processor-Sharing discipline to model the characteristics of the system when handling data traffic. Suppose that files (or bursts) arrive according to a Poisson process with parameter $\lambda_d$ files (or bursts) per unit of time. On the file-level this is a reasonable assumption. Also the amount of service requested for each file (file size) is exponentially distributed with mean $\frac{1}{\mu_d}$ kbits per file and the service discipline is PS. This queueing system can be denoted with $M/M/1-PS$. The state of the queueing process is represented by the number of customers in service.

The total capacity of the system is $C_{total}$, which is divided between the voice and data customers. Voice customers require capacity to be allocated in a circuit-switched manner. When allocating capacity voice customers take capacity from the total capacity for a certain time duration. The capacity left over after allocating capacity for voice customers can be used by the data customers. The capacity allocated for voice customers is defined as $C_v(i)$, when there are $i$ voice customers present. Also we define the capacity of the server for data, when there are $i$ voice users present in the system, as $C_d(i)$. The capacity for data customers fluctuates in time. This is because the number of

voice customers in the system is a random variable.

$$C_d(i) = C_{total} - C_v(i) \tag{2.4}$$

The randomness of $C_d(i)$ imposed by the voice process can be modeled in 2-dimensional queueing system. We briefly speak about this in section 2.6. In this thesis we will decouple the 2-dimensional system using the so called quasi-stationary regime, see section 3.1.

Under the quasi-stationary regime the number of ongoing voice calls varies much slower than the number of data customers. The data process reaches a statistical equilibrium in the time interval between a change in the number of ongoing voice calls. A single individual data customer finds upon arrival a system in equilibrium with $i$ voice customers present with possibility $p_v(i)$, where $p_v(i)$ is the possibility of the voice process being in state $i$. We look at the data process in each state of the voice process and assume that data customers hardly see any changes in the number of ongoing voice calls during their stay. Because for a single individual data customer the capacity used for voice calls is fixed at $C_v(i)$ during its stay, the capacity for this period is fixed at $C_d(i)$ under the quasi-stationary regime. The time randomness of $C_d(i)$ is in this way nearly gone. We want to know the mean conditional sojourn time of a data customer when the capacity for data is fixed at $C_d(i)$. In the quasi-stationary regime we take the weighted sum over all states of the voice process. In this way the effect of the random capacity imposed by the voice process is taken into account.

We model the data traffic process with a $M/M/1$ queue with the 'Processor-Sharing' (PS) discipline. From the balance equations one can derive the steady-state probabilities.

$$p_d(j-1)\lambda_d = p_d(j)C_d(i)\mu_d \qquad\qquad j = 1, 2, \cdots, \infty.$$

Together with the normalization condition $\sum_{j=0}^{\infty} p_d(j) = 1$ and the stability condition $\rho_d(i) = \frac{\lambda_d}{C_d(i)\mu_d} < 1$, the following steady-state probabilities are derived.

$$p_d(j) = (\lambda_d / C_d(i)\mu_d)^j p_d(0) = \rho_d(i)^j(1-\rho_d(i)) \qquad j = 0, 1, \cdots, \infty, \tag{2.5}$$

The average number of files in the system is

$$E[N_d] = \sum_{j=0}^{\infty} j \cdot p_d(j) = \frac{\rho_d(i)}{1-\rho_d(i)}, \tag{2.6}$$

and using *Little's theorem*, see [14], $E[N_d] = \lambda_d \cdot E[V]$, the average time in the system for each file is

$$E[V] = \frac{1/\mu_d}{C_d(i)(1-\rho_d(i))}. \tag{2.7}$$

The PS queue also has an insensitivity property. The distribution of the queue length in equation (2.5) depends only on the *mean* of the distribution of

the file size. The following formula, see [12, p. 168], shows that the average time spent in the system is proportional to the file size $\tau$:

$$E[V(\tau)] = \frac{\tau}{C_d(i)(1-\rho_d(i))}, \tag{2.8}$$

which readily implies (2.7).

There is a result for the variance of the sojourn time conditional on the service requirement as referred to in [17, p. 142] for a general service requirement distribution. First let us introduce some notation. Here $B$ and $T$ are distributed as the service requirement in kbits and the total amount of work in kbits at an arbitrary moment in time. $W$ is distributed as the (steady-state) waiting time in seconds in a $M/G/1-FCFS$ queue, see [5, expression 4.82] and $C$ is the service speed in kbits per second.

$$P\{W \leq t\} = (1-\rho_d(i)) \sum_{n=0}^{\infty} \rho_d^n(i) [\frac{1}{E[B]} \int_{x=0}^{t} P\{B > x\} \mathrm{d}x]^{n\star} \tag{2.9}$$

Here, the symbol $\star$ denotes the convolution operator for probability distributions, i.e., for a distribution function $H(x)$, $x \geq 0$, we define $H(x)^{0\star} := 1$, for all $x \geq 0$, and for $n \in N$ and $x \geq 0$,

$$H(x)^{(n+1)\star} := \int_{u=0}^{x} H(x-u)^{n\star} \mathrm{d}H(u).$$

In particular, $H(x)^{1\star} = H(x)$, $x \geq 0$.

Equation (2.9) can be interpreted in the following manner. Here $B$, $T$ and $W$ are as in equation (2.9). $FCFS$ and $PS$ refer to the service discipline of the system. As can be seen from equation (2.10) $W$ and $T$ are independent of the service discipline. This is a consequence of the $PASTA$ (Poisson Arrivals See Time Averages) property, see [20]. Also $C$ is the service speed in kbit per second.

$$P\{W^{FCFS} \leq t\} = P\{T^{FCFS} \leq Ct\} = P\{T^{PS} \leq Ct\} \tag{2.10}$$

$$P\{T^{PS} \leq Ct\} = \sum_{n=0}^{\infty} P\{T^{PS} \leq Ct \mid N_d^{PS} = n\} P\{N_d^{PS} = n\} \tag{2.11}$$

$$P\{N_d^{PS} = n\} = (1-\rho_d(i))\rho_d(i)^n \tag{2.12}$$

$$P\{B_i^{residual} \leq Ct\} = \frac{1}{E[B]} \int_{x=0}^{Ct} P\{B > x\} \mathrm{d}x \tag{2.13}$$

$$P\{T^{PS} \leq Ct \mid N_d^{PS} = n\} = P\{\sum_{i=1}^{n} B_i^{residual} \leq Ct\} \tag{2.14}$$

Combined, equations (2.10) through (2.14) yield (2.9). See for instance [4] for equation (2.10) and [5, p. 111] for equation (2.14).

An expression for the variance of the sojourn time conditional on the service requirement is, see [17] and [24],

$$\mathrm{Var}[V(\tau)] = \frac{2}{(1-\rho)^2}\int_{u=0}^{\tau/C}(\tau/C-u)P\{W>u\}\mathrm{d}u, \tag{2.15}$$

where $\tau$ is the file size in kbits, $C$ the service speed in kbit per second and $\rho = \frac{\lambda}{C\mu}$ is the system load.
The reader should be warned that in the following part the index for data is dropped. All parameters are for data customers. Now let us look at this variance when the service requirement is exponentially distributed with parameter $\mu$ (the average file size in kbits). If this is the case then the *residual* service requirement is also exponentially distributed with parameter $\mu$.

$$\begin{aligned}
P\{T>x\} &= \sum_{n=0}^{\infty} P\{T>x \mid N=n\}P\{N=n\} \\
&= \sum_{n=0}^{\infty} P\{T>x \mid N=n\}(1-\rho)\rho^n \\
&= \sum_{n=1}^{\infty} P\{\sum_{k=1}^{n} B_k^{residual}>x\}(1-\rho)\rho^n \\
&= \sum_{n=1}^{\infty} P\{\sum_{k=1}^{n} B_k>x\}(1-\rho)\rho^n \\
&= \sum_{n=1}^{\infty}\sum_{i=0}^{n-1} e^{-\mu x}\frac{(\mu x)^i}{i!}(1-\rho)\rho^n \\
&= (1-\rho)\sum_{i=0}^{\infty} e^{-\mu x}\frac{(\mu x)^i}{i!}\sum_{n=i+1}^{\infty}\rho^n \\
&= e^{-\mu x}\sum_{i=0}^{\infty}\frac{(\mu x)^i}{i!}\rho^{i+1} \\
&= \rho e^{-\mu x}\sum_{i=0}^{\infty}\frac{(\rho\mu x)^i}{i!} \\
&= \rho e^{-(1-\rho)\mu x}.
\end{aligned}$$

Now if $B$ is exponentially distributed with parameter $\mu$, $1/\mu$ in kbits per file, capacity $C$ in kbits per second and $\rho = \frac{\lambda}{C\mu}$ the load, then

$$\begin{aligned}
\mathrm{Var}[V(\tau)] &= \frac{2}{(1-\rho)^2}\int_{u=0}^{\tau/C}(\tau/C-u)P\{W>u\}\mathrm{d}u \\
&= \frac{2}{(1-\rho)^2}\int_{u=0}^{\tau/C}(\tau/C-u)\rho e^{-(1-\rho)\mu C u}\mathrm{d}u \\
&= \frac{2\rho}{C^2(1-\rho)^4\mu^2}[e^{-(1-\rho)\mu\tau}+(1-\rho)\mu\tau-1].
\end{aligned}$$

Note that this function is almost linear in $\tau$, when $\tau$ is large. This result gives the variance conditional on the file size when the service speed is fixed at $C$.

## 2.5 Throughput and goodput

One of the QoS constraints for data traffic is the goodput per user. The goodput per user is the number of bits a user wants to send divided by the time it takes to complete the user's request. Goodput is related to throughput in the following manner:

$$\text{Goodput} = \text{Throughput} \cdot \frac{\tau_{user}}{\tau_{actual}}. \tag{2.16}$$

Here $\tau_{user}$ is the number of bits the user wants to send and $\tau_{actual}$ is the number of bits that are actually sent over the channel. There is a difference between the two because of retransmissions and overhead generated by several layers of the network. In this thesis we will focus on throughput as one of the QoS constraints. This is because $\tau_{user}$ is known, while it is not clear what $\tau_{actual}$ will be due to uncertainty about the number of bits used for overhead and the number of retransmissions in the network.
The definition of throughput seen by a user is

$$TH^{user}(\tau) = \frac{\tau}{E[V(\tau)]}, \tag{2.17}$$

where $V(\tau)$ is the sojourn time conditional on the service requirement. This is the throughput on the user level, not on the system level. On the system level we look at the total number of bits that leave the system per unit of time, while on the user level we look at the time it takes to send $\tau$ bits for a given user (when there are no retransmissions and there is no overhead). Let $TH_i^{user}(\tau)$ be the throughput on the user level when there are $i$ voice users in the system and the user wants to send $\tau$ kbits.

$$TH_i^{user}(\tau) = \frac{\tau}{E[V(\tau)]_i} = \frac{\tau}{\frac{\tau}{C_d(i)(1-\rho_d(i))}} = C_d(i) \cdot (1 - \rho_d), \tag{2.18}$$

Now let $TH_i^{system}$ be the system throughput of the data traffic when there are $i$ voice users in the system and the offered load is such that the system is stable ($\rho_d(i) < 1$).

$$TH_i^{system} = (1 - \rho_d(i)) \cdot 0 + \rho_d(i) \cdot C_d(i) = \frac{\lambda_d}{\mu_d} \tag{2.19}$$

If the offered load is too much for the system ($\rho_d(i) \geq 1$), then the system would never be empty and the throughput on the system level would be the same as the maximum service speed $C_d(i)$. There is a relation between the throughput on the system and user level in a stable system.

$$E[N_d] \cdot TH_i^{user} = \frac{\rho_d(i)}{1 - \rho_d(i)} \cdot C_d(i)(1 - \rho_d(i)) = \frac{\lambda_d}{\mu_d} = TH_i^{system},$$

$$TH_i^{user} = \frac{TH_i^{system}}{E[N_d]}, \tag{2.20}$$

where $N_d$ is the number of data users in the system. The PS discipline is fair in the sense that it equally divides the system throughput among the users.

## 2.6 The integrated model

Until now we have looked at the voice-call process and the data-call process in isolation. In the real system, both call types are integrated, and share the total capacity. The sharing of the systems capacity causes the two processes to interact with each other. The two processes can be modeled by using a 2-dimensional Markov queueing system. The state space of this queueing system can be denoted by $(V, D)$, where $V$ is the number of voice calls and $D$ is the number of data calls. Here, the voice call QoS constraint again requires the voice calls to be handled in a circuit-switched fashion. Because of the integration of both call-types, capacity is shared between the types. If the total capacity (service rate) of the system is $C_{total}$ and the service rate per call is $r_v$ and $r_d$, then the following restriction should be met:

$$V \cdot r_v + D \cdot r_d \leq C_{total} \tag{2.21}$$

In equation (2.21) again the capacity is shared between data and voice traffic, as in equation (2.4). The main difference between the two equations is the way capacity is allocated. In equation (2.4) the capacity for data traffic is the left over capacity of the total capacity after using capacity for voice traffic. The sharing of capacity in equation (2.21) is more direct and acts instantaneously on changes in the state of data and voice traffic. In equation (2.21) we also assume a minimal (non-zero) service rate for serving calls. With this assumption the total number of calls in the system is bounded.

In the 2-dimensional system the effect of randomness of the capacity available for data customers is instantaneously present. Each arrival or departure of a voice customer allocates or frees a part of $C_{total}$. Data customers can only use that part of the total capacity left over by voice customers. In section 2.4 we stated the following in equation (2.4) :

$$C_d(i) = C_{total} - C_v(i)$$

This means that the capacity that can be used by data customers depends on the number of voice customers present in the system. In the 2-dimensional model $C_d(i)$ fluctuates heavily. Still the 2-dimensional model captures the dynamics of the integrated voice and data process in a more detailed way. We are interested in the mean conditional sojourn time of data customers and the maximal allowed voice load under QoS constraints. These results should be obtained using small amounts of computing time and the analysis of the models should be tractable.

In chapter 2 of [17] the integrated 2-dimensional model is discussed. The author found a solution to the queue length problem through use of a matrix geometric approach. The mathematical features of the integrated 2-dimensional model are harder to analyse than those of the $M/G/c_v/c_v$ and $M/G/1/c_v - PS$ queue. We want our model to be such that it is easier to analyse. For this reason we will use a different approach towards integrating the voice and data process. For PS queue in isolation there is a linearity result for the mean conditional sojourn time. Some of the nice properties that hold for the PS

model in isolation do not hold in the integrated model. Harder mathematical models often require more computing power. We also want that the result of our model can be computed using less computing time. For these reasons we will look for another way to describe the model, while still taking into account the interaction between the two processes.

Because of the differences in handling between circuit-switched and packet-switched calls and the time scales on which the processes evolve, one could say that data calls see voice calls as nearly stationary during their stay in the system. We will use this to model the system characteristics using the so called quasi-stationary regime in the next chapter.

# Chapter 3

# The quasi-stationary regime and admission controlled transmission delay

In this chapter we will introduce the quasi-stationary regime. This regime will give the possibility to model the system using several 1-dimensional queueing systems for the data process. In section 2.6 we argued that a 2-dimensional model is more precise but harder to handle mathematically. We want our model to be less complex and results need to be computed without a lot of computing power or computing time. Because of the linearity result for the mean conditional sojourn times in the PS queue both features are present in the quasi-stationary regime model. Still we need our model to cope with admission control. Admission control ensures that there is a maximum number of customers simultaneously present. In the previous chapter all models for packet-switched systems where such that they accepted infinitely many data customers simultaneously. Accepting all customers could lead to congestion in the system. For this reason there is a limit on the number of data customers present in the system. We will discuss the consequences for the models when the number of simultaneous data calls is limited. Admission control is necessary for the reason that in some states of the voice process the data load can exceed the available capacity left over by the voice users. The admission control protocol can reject calls in order to prevent that too many calls enter the system. In this way it is possible to ensure that the system will never become instable. If no admission control is applied, the system could become instable.

We assume that data traffic has a minimal $r_-$ and a maximal processing rate $r_+$. The maximum number of data customers in the system is limited by the hardware of the system. For now we will assume that the maximum number of data customers simultaneously present in the system is fixed at $c_d$. The steady-state probabilities for this system are given by

$$p_d(j) = (\frac{\lambda_d}{C_d(i) \cdot \mu_d})^j \cdot p_d(0) \tag{3.1}$$

$$= (\frac{\lambda_d}{C_d(i) \cdot \mu_d})^j \cdot \frac{1}{\sum_{k=0}^{c_d} p_d(k)} \tag{3.2}$$

$$= \frac{1-\rho_d(i)}{1-\rho_d^{c_d+1}(i)} \cdot \rho_d^j(i), \tag{3.3}$$

where $\rho_d(i) = \frac{\lambda_d}{C_d(i)\cdot\mu_d}$ is the relative data load when there are $i$ voice users in the system.

In section 3.1 we introduce the quasi-stationary regime. This regime is used later on to simplify the integrated model. The opposite of the quasi-stationary regime is the fluid regime. We will briefly look into this regime, but the focus of this thesis is on the quasi-stationary regime. Next, we will look at the consequences of admission control for the PS model. First we will solve the mean conditional sojourn time by using differential equations for the $M/M/1/c_d - PS$ model. This is done in section 3.2. Because we want the service requirements to be generally distributed we will extend the model. In section 3.3 we state the system of differential equations for the mean conditional sojourn time in the $M/G/1/c_d - PS$ model and in section 3.4 we will solve these equations using Laplace transforms in the case that there can be only two customers present simultaneously. This again results in a linear form for the mean conditional sojourn time. The system of differential equations is not easy to solve in the case of generally distributed service requirements and still takes a lot of computing power for large numbers of $c_d$. We will turn to another approach. In section 3.5 we make use of the characteristics of symmetric queues. This leads to a solution for the mean conditional sojourn time problem that is of a simple linear form. With this we have a satisfactory result for the dimensioning of the system.

## 3.1 Quasi-stationary and fluid regimes

One call type, say voice calls, could occupy the system for a significantly longer time than another call type, say data calls. We introduce two limit regimes to make approximations for the mean conditional sojourn time of the data calls. These limit regimes are the fluid regime and the quasi-stationary regime. Under the fluid regime the voice process evolves on infinitely fast time scale and under the quasi-stationary regime the voice process evolves on a infinitely slow time scale, see [3].

Under the quasi-stationary regime the capacity that can be used to process the data calls will change infinitely slow, so it is possible to reduce the effects of the random capacity environment imposed by the voice call process. This approach in which we will look at the data process with a near constant number of voice users is called the ***quasi-stationary regime*** (see [6]).
The following part is taken from the CWI website of Matthieu Jonckheere, see [9]: A quasi-stationary regime is a scaled version of the original process, constructed by making the speeds of the different classes tend successively towards infinity in a certain order, while keeping the incoming load of each class constant. In other words, the dynamics of each class are infinitely increased in a given order. Hence, in the limiting regimes, each class sees the dynamics of other classes as infinitely slow or infinitely fast depending on the order of the successive limits. (The limiting regimes obtained are insensitive to service

time distributions except for their means and have a nice product-form stationary measure.) With certain monotonicity properties of the transitions, the stability region of the initial process might be bounded in terms of the stability region of these limiting processes, which can eventually be computed more easily. It is also tempting to think that the performance of the limiting quasi-stationary regime constitutes worst or best cases and can then provide bounds of the performance for the original network. This has been observed for certain monotonic systems but turns out to be difficult to prove since these bounds are not sample-path bounds.

To examine the data process with queueing theory we require that the process is in equilibrium. Under the quasi-stationary regime data customers see the number of voice customers as a constant during their stay in the system. Because the data process evolves on a time-scale that is much faster than that of the voice process we could assume that the data process reaches a statistical equilibrium before a change in the voice process occurs. An single individual arriving data customer could enter the system in a state of equilibrium with $i$ voice customers present with probability $p_v(i)$. Still the voice process changes its state, but on a slower time-scale than the data process. To cope with the effects of a random environment imposed by the voice process we introduce the quasi-stationary regime. We model the data process in isolation and change the capacity for the data process according to the state of the voice process. The capacity of the data process obeys equation (2.4) and so depends on the state of the voice process.

To combine voice and data traffic in one model we compute the sojourn time of the data customers in a queue in isolation with capacity $C_d(i)$ and take the weighted sum over all states of the voice process. The voice process is also analyzed in isolation using a $M/G/c_v/c_v$ queueing model. We can compute the steady-state probability that the system is in a state where there are $i$ voice calls active. This probability is equal to the percentage of time spent in that specific state with $i$ ongoing voice calls. By taking the weighted sum over all states of the voice process one can make approximate statements about the mean conditional sojourn time of data customers with randomly varying capacity imposed by the voice process.

We will now show how to compute the average number of data calls in the system under the quasi-stationary regime. Let $N_d$ be the number of data calls in the system, $c_v$ the maximum number of simultaneous voice calls and $E[N_d]_i$ the average number of data calls in the presence of a constant number $i$ of voice calls. Also denote the steady-state probabilities of the voice call process by $p_v(i)$. Let $\rho_d(i) = \frac{\lambda_d}{C_d(i)\mu_d}$ be the data traffic load, where $C_d(i)$ is the capacity that can be used for data calls when there are $i$ voice calls present. For now assume that in all states of the voice process the offered data load is such that the system is stable ($\rho_d(i) < 1 \ \forall i$). Now, let us take a look at the average number of data calls in the system under the quasi-stationary regime.

$$E[N_d]_{qs} = \sum_{i=0}^{c_v} p_v(i) \cdot E[N_d]_i \tag{3.4}$$

$$= \sum_{i=0}^{c_v} p_v(i) \cdot \frac{\rho_d(i)}{1-\rho_d(i)} \tag{3.5}$$

$$= \sum_{i=0}^{c_v} p_v(i) \cdot \frac{\frac{\lambda_d}{\mu_d}}{C_d(i) - \frac{\lambda_d}{\mu_d}}. \tag{3.6}$$

Another approximation is the ***fluid regime*** in which the time scale on which the voice process evolves is infinitely faster than that of the data process, see [3]. The capacity available for data calls in all different states of the voice process is taken as the average capacity over all different states of the voice process, so $\overline{C_d} = \sum_{i=0}^{c_v} p_v(i) C_d(i)$. Thus in the fluid regime $C_d(i)$ is taken to be $\overline{C_d}$ in all different states of the voice call process.

$$E[N_d]_{fl} = \frac{\overline{\rho_d}}{1-\overline{\rho_d}} \tag{3.7}$$

$$= \frac{\frac{\lambda_d}{\mu_d}}{\overline{C_d} - \frac{\lambda_d}{\mu_d}}. \tag{3.8}$$

The throughput on the system level using the fluid approximation is

$$TH_{fl}^{system} = \sum_{i=0}^{c_v} p_v(i) \cdot TH_i^{system} \tag{3.9}$$

$$= \frac{\lambda_d}{\mu_d}. \tag{3.10}$$

The throughput on the user level using the fluid approximation is

$$TH_{fl}^{user} = \sum_{i=0}^{c_v} p_v(i) \cdot TH_i^{user} \tag{3.11}$$

$$= \overline{C_d} - \frac{\lambda_d}{\mu_d}. \tag{3.12}$$

Also, the quasi-stationary regime can be used to approximate the throughput

$$TH_{qs}^{user} = \sum_{i=0}^{c_v} p_v(i) \cdot TH_i^{user} \tag{3.13}$$

$$= \sum_{i=0}^{c_v} p_v(i) \cdot (C_d(i) - \frac{\lambda_d}{\mu_d}). \tag{3.14}$$

The sojourn time conditional on the amount of work can also be computed under the quasi-stationary regime.

$$E[V(\tau)]_{qs} = \sum_{i=0}^{c_v} p_v(i) \cdot E[V(\tau)]_i \tag{3.15}$$

$$= \sum_{i=0}^{c_v} p_v(i) \cdot \frac{\tau}{C_d(i) - \frac{\lambda_d}{\mu_d}}, \tag{3.16}$$

$$E[V(\tau)]_{fl} = \frac{\tau}{\overline{C_d} - \frac{\lambda_d}{\mu_d}}.$$

In the previous paragraph the throughput on the user level is defined as the amount of work that needs to be processed by the system divided by the time it takes to complete the request. But as can be seen under the quasi-stationary regime it is possible to compute two different measures of throughput.

$$TH_{qs}^{user} \neq \frac{\tau}{E[V(\tau)]_{qs}}.$$

The first one is computed directly from the throughput when there are $i$ voice users, see (3.14). The other one is computed through the quasi-stationary delay given the amount of work that needs to be processed, see (3.16). If $C_d(i)$ is close to $\frac{\lambda_d}{\mu_d}$, then $E[V(\tau)]_{qs}$ will go to infinity and $\frac{\tau}{E[V(\tau)]_{qs}}$ approaches zero. So a nearly instable state affects the throughput greatly in this case, while when $TH_{qs}^{user}$ is computed through $TH_i^{user}$ a nearly instable state will not affect the throughput that dramatically. On the other hand one could still revert to the definition given above, which states that the throughput is calculated by dividing the total amount of work by the time it took to process the work (under the quasi-stationary regime). Because we have no preference for one of the two measures, we will use both of them.

The variance of the sojourn time under the quasi-stationary regime can also be computed via the second moment of the sojourn time. The second moment of the sojourn time when there are $i$ voice users in the system is

$$\begin{aligned} E[V(\tau)^2]_i &= \mathrm{Var}[V(\tau)]_i + (E[V(\tau)]_i)^2 \\ &= \frac{2}{(1-\rho_d(i))^2}\left(\int_{u=0}^{\tau/C_d(i)} \left(\frac{\tau}{C_d(i)} - u\right) P\{W > u\}\mathrm{d}u + \frac{1}{2}\left(\frac{\tau}{C_d(i)}\right)^2\right). \end{aligned}$$

Now the quasi-stationary variance can be computed.

$$E[V(\tau)^2]_{qs} = \sum_i p_v(i) \cdot E[V(\tau)^2]_i, \tag{3.17}$$

$$\mathrm{Var}[V(\tau)]_{qs} = E[V(\tau)^2]_{qs} - (E[V(\tau)]_{qs})^2. \tag{3.18}$$

If the service requirement is exponentially distributed with parameter $\mu_d$ then

$$\begin{aligned} \mathrm{Var}[V(\tau)]_{qs} &= \sum_i p_v(i) \cdot \Big(\frac{2\rho_d(i)}{C_d^2(1-\rho_d(i))^4\mu_d^2} \cdot [e^{-(1-\rho_d(i))\mu_d\tau} + (1-\rho_d(i))\mu_d\tau - 1 \\ &\quad - \frac{(1-\rho_d(i))^2\mu_d^2}{2\rho_d(i)} \cdot \tau^2]\Big) - \Big(\sum_i p_v(i) \cdot \frac{\tau}{C_d(i) - \frac{\lambda_d}{\mu_d}}\Big)^2. \end{aligned}$$

It is now possible to compute the QoS constraints for data traffic under the quasi-stationary regime with the measures presented in this section. However, in some states of the voice call process, the capacity left over for data calls could be such that the offered load exceeds one. This is called a local instability, and can occur even in the case of local stability, i.e., $\frac{\lambda_d}{\mu_d} < \sum_{i=0}^{c_v} p_v(i)C_d(i)$. When the system is in a state of local instability ($\rho_d(i) \geq 1$) the performance measures are not valid. In the next section we will introduce admission control to ensure uniform stability.

## 3.2 First moment of $V(\tau)$ in a $M/M/1/c-PS$ queueing system

We are interested in the sojourn time of an accepted customer in a $M/M/1/c-PS$ system. This system has Poisson arrivals for which the $PASTA$ property (see [20]) holds. If we assume that the service requirements of the customers are exponentially distributed, it is possible to obtain this sojourn time through a set of differential equations. We will derive these equations in the following. Here $V_n(\tau)$ is the sojourn time in seconds of an arriving data customer when there are $n$ other customers in the system upon arrival and the arriving customer has a service requirement of $\tau$ kbit.
For $1 \leq n \leq c_d - 2$

$$\begin{aligned} E[V_n(\tau+\Delta)] = \lambda_{n+1} \cdot (n+1) \cdot \frac{\Delta}{r_{n+1}} \cdot E[V_{n+1}(\tau) + (n+1) \cdot \frac{\Delta}{r_{n+1}}] \\ +\mu \cdot r_{n+1} \cdot \frac{n}{n+1} \cdot (n+1) \cdot \frac{\Delta}{r_{n+1}} \cdot E[V_{n-1}(\tau) + (n+1) \cdot \frac{\Delta}{r_{n+1}}] \\ +(1 - \lambda_{n+1} \cdot (n+1) \cdot \frac{\Delta}{r_{n+1}} - \mu \cdot r_{n+1} \cdot \frac{n}{n+1} \cdot (n+1) \cdot \frac{\Delta}{r_{n+1}}) \\ \cdot E[V_n(\tau) + (n+1) \cdot \frac{\Delta}{r_{n+1}}] \end{aligned}$$

During the processing of $\Delta$ kbit of our test customer, when there are $n$ other customers in the system and the service discipline is PS, every other customer gets an equal share of processing time. So every customer is served by the processor for $\frac{\Delta}{r_{n+1}}$ seconds. Now if the test customer has received $\frac{\Delta}{r_{n+1}}$ of processor time, so will all the other customers. The total time to process $\Delta$ kbit for all customers is equal to $\frac{(n+1)\Delta}{r_{n+1}}$. The intensity of an arrival during the time interval $\frac{(n+1)\Delta}{r_{n+1}}$ is $\lambda_{n+1} \cdot \frac{(n+1)\cdot\Delta}{r_{n+1}}$. After processing $\Delta$ kbit of all customers and having an arrival during the processing the remaining sojourn time can be expressed through $E[V_{n+1}(\tau) + (n+1) \cdot \frac{\Delta}{r_{n+1}}]$. Multiply this fraction with the arrival intensity to get the first part of the equation. This reasoning can be applied to departures and the event that nothing happens during the time interval $(n+1) \cdot \frac{\Delta}{r_{n+1}}$. Now by rearranging the terms, dividing by $\Delta$ and taking the limit of $\Delta$ to zero, we get a set of differential equations.
For $1 \leq n \leq c_d - 2$

$$\begin{aligned} \lim_{\Delta \downarrow 0} \frac{E[V_n(\tau+\Delta)] - E[V_n(\tau)]}{\Delta} = \frac{n+1}{r_{n+1}} \\ +\lambda_{n+1} \cdot \frac{n+1}{r_{n+1}} \cdot (E[V_{n+1}(\tau)] - E[V_n(\tau)]) \\ +\mu \cdot n \cdot (E[V_{n-1}(\tau)] - E[V_n(\tau)]) \end{aligned}$$

For $n = 0$,

$$\lim_{\Delta \downarrow 0} \frac{E[V_0(\tau+\Delta)] - E[V_0(\tau)]}{\Delta} = \frac{1}{r_1} + \frac{\lambda}{r_1} \cdot (E[V_1(\tau)] - E[V_0(\tau)])$$

For $n = c_d - 1$,

$$\lim_{\Delta \downarrow 0} \frac{E[V_{c_d-1}(\tau+\Delta)] - E[V_{c_d-1}(\tau)]}{\Delta} = \frac{c_d}{r_{c_d}} + \mu \cdot c_d \cdot (E[V_{c_d-2}(\tau)] - E[V_{c_d-1}(\tau)])$$

This together gives a set of linear differential equations of the following form:

$$\begin{pmatrix} E[V_0(\tau)]' \\ \vdots \\ \\ \\ \vdots \\ E[V_{c_d-1}(\tau)]' \end{pmatrix} = A \cdot \begin{pmatrix} E[V_0(\tau)] \\ \vdots \\ \\ \\ \vdots \\ E[V_{c_d-1}(\tau)] \end{pmatrix} + b \tag{3.19}$$

Here $A$ is a tri-diagonal matrix of dimensions $c_d \times c_d$ with elements $A_{i,i-1} = (i-1) \cdot \mu$, $A_{i,i+1} = \frac{i \cdot \lambda}{r_i}$ and $A_{i,i} = -\sum_{j \neq i} A_{i,j}$:

$$\begin{pmatrix} -\frac{\lambda}{r_1} & \frac{\lambda}{r_1} & 0 & \dots & & \dots & 0 \\ \mu & -2\frac{\lambda}{r_2} - \mu & 2 \cdot \frac{\lambda}{r_2} & 0 & \dots & \dots & 0 \\ & \ddots & \ddots & \ddots & & & \\ \\ 0 & & \dots & 0 & (c_d-2)\cdot\mu & -(c_d-1)\frac{\lambda}{r_{c_d-1}} - (c_d-2)\cdot\mu & (c_d-1)\frac{\lambda}{r_{c_d-1}} \\ 0 & & & \dots & 0 & (c_d-1)\cdot\mu & -(c_d-1)\cdot\mu \end{pmatrix}$$

and $b$ a vector of dimensions $c_d \times 1$:

$$b = \begin{pmatrix} \frac{1}{r_1} \\ \frac{2}{r_2} \\ \vdots \\ \frac{c_d-1}{r_{c_d-1}} \\ \frac{c_d}{r_{c_d}} \end{pmatrix}$$

The boundary conditions are $E[V_n(0)] = 0, \forall n$, since when a customer has no service requirement it will leave the system immediately.

This system of differential equations can be solved following this procedure:

- compute the $c_d$ eigenvalues $v_i$ and $c_d$ eigenvectors $w_i$ of $A$.
- try a solution of the form

$$\begin{pmatrix} E[V_0(\tau)] \\ \vdots \\ \\ \\ \vdots \\ E[V_{c_d-1}(\tau)] \end{pmatrix} = C_0 \tau w_0 + \sum_{i=1}^{c_d-1} C_i e^{v_i \tau} w_i + \begin{pmatrix} a_0 \\ \vdots \\ \\ \\ \vdots \\ a_{c_d-1} \end{pmatrix} \tag{3.20}$$

- use the boundary conditions $E[V_i(0)] = 0$ and the equations of the system of differential equations itself to find the constants $C_i$ and $a_i$.

The solution form which is used contains a linear term of $\tau$. One of the eigenvalues ($v_0$ with eigenvector $w_0$) of $A$ will be zero, because of its structure. For this reason one needs to add a linear term of $\tau$ to the solution form. With the solution form, the derivative of the solution form, the boundary conditions and the system equations one can form enough linear equations to solve the remaining undetermined constants $C_i$ and $a_i$.

Finding the eigenvalues of the matrix $A$ involves finding the roots of $Det(A-\lambda I) = 0$. There are exact solutions for dimensions below 5, but for dimensions larger than or equal to 5 there are generally no exact solutions and one has to resort to numerical methods to find them approximately, see [21]. Solving the eigenvalue problem numerically can give rounding errors and other numerical issues.

We will solve the system of differential equations for the case where only two customers can be present simultaneously. To get numerical results we will set $\lambda = 2$, $\mu = 4$, $r_1 = r_2 = 1$. This results in the following system of differential equations:

$$\begin{pmatrix} E[V_0(\tau)]' \\ E[V_1(\tau)]' \end{pmatrix} = \begin{pmatrix} -2 & 2 \\ 4 & -4 \end{pmatrix} \cdot \begin{pmatrix} E[V_0(\tau)] \\ E[V_1(\tau)] \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

This system can be solved following this procedure:

- compute the eigenvalues $v_1$, $v_2$ and eigenvectors $w_1$, $w_2$ of $A$;
- try a solution of the form $\begin{pmatrix} E[V_0(\tau)] \\ E[V_1(\tau)] \end{pmatrix} = C_1 e^{v_1 \tau} w_1 + C_2 \tau w_2 + \begin{pmatrix} a \\ b \end{pmatrix}$
- use the boundary conditions $E[V_i(0)] = 0$ and the equations of the system of differential equations itself to find $C_1$, $C_2$, $a$ and $b$.

We will follow the procedure in the example:

The eigenvalues and eigenvectors of $A$ are

- $v_1 = -6$ with $w_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$
- $v_2 = 0$ with $w_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Now try a solution of the form

$$\begin{pmatrix} E[V_0(\tau)] \\ E[V_1(\tau)] \end{pmatrix} = C_1 e^{-6\tau} \begin{pmatrix} 1 \\ -2 \end{pmatrix} + C_2 \tau \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}$$

The boundary conditions are $E[V_0(0)] = 0$ and $E[V_1(0)] = 0$. Use this in the solution to obtain

$$\begin{aligned} C_1 + a &= 0 \\ -2C_1 + b &= 0 \end{aligned}$$

Now set equations of the system of differential equations equal to the differentiated solution

$$\begin{aligned} -2E[V_0\tau] + 2E[V_1\tau] + 1 &= E[V_0(\tau)]' = -6C_1e^{-6\tau} + C_2 \\ 4E[V_0\tau] - 4E[V_1\tau] + 2 &= E[V_1(\tau)]' = 12C_1e^{-6\tau} + C_2 \end{aligned}$$

which leads to

$$\begin{aligned} -6C_1e^{-6\tau} - 2a + 2b + 1 &= E[V_0(\tau)]' = -6C_1e^{-6\tau} + C_2 \\ 12C_1e^{-6\tau} + 4a - 4b + 2 &= E[V_1(\tau)]' = 12C_1e^{-6\tau} + C_2 \end{aligned}$$

Notice that $-2a + 2b + 1 = C_2 = 4a - 4b + 2$. Now substitute $a$ and $b$ with $C_1$ according to the boundary conditions and solve the equation for $C_1$. This results in $C_1 = \frac{1}{18}$ and $C_2 = \frac{4}{3}$. Now again use the boundary conditions and the solution with $C_1$ and $C_2$ filled in to obtain $a = -\frac{1}{18}$ and $b = \frac{1}{9}$.

The solution to this system of differential equations is:

$$\begin{pmatrix} E[V_0(\tau)] \\ E[V_1(\tau)] \end{pmatrix} = \frac{1}{18}e^{-6\tau}\begin{pmatrix} 1 \\ -2 \end{pmatrix} + \frac{4}{3}\tau\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} -\frac{1}{18} \\ \frac{1}{9} \end{pmatrix}$$

Notice that there is an asymptotic linearity in $\tau$. If $\tau$ is large, $e^{-6\tau}$ will be near zero.
If one wants to obtain the mean conditional sojourn time for an arbitrary customer, one needs to average over all states of the data process the arriving customer can see when entering the system.

$$E[V(\tau)] = \frac{p_0}{p_0 + p_1} \cdot E[V_0(\tau)] + \frac{p_1}{p_0 + p_1} \cdot E[V_1(\tau)] = \frac{4}{3}\tau$$

Now we see that for an arbitrary customer the mean conditional sojourn time will be linear in $\tau$. Arriving in a system with $n$ other data customers gives a little bit more information on the mean conditional sojourn time. For large $\tau$ this difference will be relatively small. For the solution of the system of differential equations one can see that arriving in an empty system will give a slightly smaller sojourn time then when there is already a customer present. When $\tau$ gets larger this advantage will not be significant compared to the impact of the linear term of $\tau$ in the solution. The difference between entering an empty system and entering a system with one customer present will be $\frac{3}{18}$ in this example when $\tau$ is very large.

## 3.3 Generally distributed service requirements

We are really interested in the first moment of the conditional sojourn time of an arriving customer for generally distributed service requirements. First, let us discuss the possible events that can occur during the processing of $\tau$ kbit of the test customer.

- Arrival of new customer: Customers arrive according to a Poisson process.

- Departure of customer: Customers depart when their remaining service requirement is zero. Their service requirement is generally distributed.

- No arrival or departure: All customers have processed $\tau$ kbit. No new customer arrives and all service requirements of other customers are strictly positive after the processing of $\tau$ kbit.

We now investigate the intensity of the different events. What is the chance that during the processing time $(n+1) \cdot \frac{\Delta}{r_{n+1}}$ an event will occur?
Because the arrival process is Poisson, the arrival intensity of new customers is $\lambda_{n+1} \cdot (n+1) \cdot \frac{\Delta}{r_{n+1}}$.

Some notation is introduced.

- $x_i$: The service requirement of customer $i$.

- $V_n(\tau, x_1, \ldots, x_n)$: The sojourn time of an arriving customer when there are $n$ other customers in the system with a residual service requirement of $x_i$ for customer $i$.

- $B(x)$: The distribution of the service requirement distribution of a customer. The service requirement is identically distributed for all customers.

By looking at the arrival moments of a new customer we can again derive a set of differential equations for $V_n(\tau, x_1, \ldots, x_n)$ from the following:

$$\begin{aligned} E[V_n(\tau+\Delta, x_1+\Delta, \ldots, x_n+\Delta)] = \\ \frac{(n+1)\Delta}{r_{n+1}} \cdot \lambda \cdot \int_{x_{n+1}=0}^{\infty} E[V_{n+1}(\tau, x_1, \ldots, x_n) + (n+1)\cdot\Delta] \mathrm{d}B(x) \\ +(1 - \frac{(n+1)\Delta}{r_{n+1}} \cdot \lambda) \cdot E[V_n(\tau, x_1, \ldots, x_n) + (n+1)\cdot\Delta] \end{aligned}$$

Also the following equation holds,

$$E[V_n(\tau, x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)] = E[V_{n-1}(\tau, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)]$$

Rearrange the terms and take the limit of $\Delta$ to 0 to get

$$(\frac{\partial}{\partial\tau}+\sum_{i=1}^{n}\frac{\partial}{\partial x_i})E[V_n(\tau,x_1,\ldots,x_n)] = \frac{n+1}{r_{n+1}} - \frac{n+1}{r_{n+1}}\cdot\lambda\cdot E[V_n(\tau,x_1,\ldots,x_n)] + \frac{n+1}{r_{n+1}}\cdot\lambda\cdot\int_{x_{n+1}=0}^{\infty} E[V_{n+1}(\tau,x_1,\ldots,x_{n+1})]\mathrm{d}B(x_{n+1})$$

This again leads to a system of differential equations. Because of generally distributed service requirements one can not use the memoryless property of the residual service requirement. This makes it a lot harder to solve these equations. In the next section we will solve the differential equations using Laplace transforms for a system which allows only two data customers simultaneously.

## 3.4 First moment of $V(\tau)$ in a $M/G/1/2-PS$ queueing system

Another approach towards solving the differential equations is the use of Laplace transforms. For simplicity and to gain insight we will analyse the first moment of the conditional sojourn time in a system in which only two customers can be present simultaneously. Customers arriving when two customers are in service are not accepted and will be blocked. Again, a differential equation can be derived to determine $E[V_i(\tau)]$. The reader should be warned that the index $i$ is in this case the number of data customers an arriving customer finds in a system with only data customers. Also the index for data is dropped in this section, since all parameters refer to those of the data process. When there are no customers present upon arrival the only possible state change during a small quantum of service time is due to an arrival of a new customer. We will distinguish between new arrivals with less work and with more work than the residual service requirement of the customer present.

$$\begin{aligned} E[V_0(\tau+\Delta)] &= (1-\lambda\Delta)\cdot E[V_0(\tau)]+\Delta+ \\ &\quad \lambda\Delta(1-B(\tau))\cdot E[V_1(\tau,x_1\mid\tau\le x_1)] \\ &\quad +\lambda\Delta\cdot B(\tau)\cdot\int_{u=0}^{\tau}\frac{b(u)}{B(\tau)}\cdot E[V_1(\tau,u\mid\tau>u)]\mathrm{d}u \end{aligned}$$

Rearranging the terms and taking the limit of $\Delta$ to 0 yields:

$$\begin{aligned} \lim_{\Delta\downarrow 0}\frac{E[V_0(\tau+\Delta)]-E[V_0(\tau)]}{\Delta} &= -\lambda E[V_0(\tau)]+1+ \\ &\quad \lambda(1-B(\tau))\cdot 2\tau \\ &\quad +\lambda\int_{u=0}^{\tau}2u+E[V_0(\tau-u)]\mathrm{d}B(u) \end{aligned}$$

Also $E[V_i(0)]=0$.
If the residual service requirement of the customer present is more than that of the new customer, the new arrival will share the system capacity with the other customer during his entire stay. For simplicity we have taken the service-rate to be 1 kbit/sec, so the time it takes to serve the test customer is equal

to two times the amount of required service (in kbit). If the residual service requirement of the other customer is smaller than the amount of service the test customer requires, then that customer will leave during the service of the test customer. This together results in:

$$E[V_1(\tau, x_1)] = 2\tau, \text{if } \tau \leq x_1$$

$$E[V_1(\tau, x_1)] = 2x_1 + E[V_0(\tau - x_1)], \text{if } \tau > x_1$$

It is possible to transform this differential equation through Laplace transformation. Let $Q_0^\star(s)$ be the Laplace transform of $E[V_0(\tau)]$, so
$Q_0^\star(s) = \int_{\tau=0}^{\infty} e^{-s\tau} \cdot E[V_0(\tau)]\mathrm{d}\tau$.
Now

$$sQ_0^\star(s) = -\lambda Q_0^\star(s) + \frac{1}{s} + 2\lambda \cdot H^\star(s) + \lambda Q_0^\star(s) \cdot b^\star(s),$$

where

$$\begin{aligned}
H^\star(s) &:= \int_{\tau=o}^{\infty} e^{-s\tau}[\tau(1 - B(\tau)) + \int_{u=0}^{\tau} u \cdot \mathrm{d}B(u)]\mathrm{d}\tau \\
&= \int_{\tau=0}^{\infty} e^{-s\tau}(\int_{u=\tau}^{\infty} \tau \cdot \mathrm{d}B(u) + \int_{u=0}^{\tau} u \cdot \mathrm{d}B(u))\mathrm{d}\tau \\
&= \int_{\tau=0}^{\infty} e^{-s\tau} E[min(\tau, B)]\mathrm{d}\tau \\
&= \int_{x=0}^{\infty} \frac{1}{s^2}[1 - e^{-sx}]\mathrm{d}B(x) \\
&= \frac{1}{s^2}\int_{x=0}^{\infty} b(x)\mathrm{d}x - \frac{1}{s^2}\int_{x=0}^{\infty} e^{-sx} b(x)\mathrm{d}x \\
&= \frac{1}{s^2}(1 - sB^\star(s))
\end{aligned}$$

$$\begin{aligned}
\int_{\tau=0}^{\infty} e^{-s\tau} \int_{u=0}^{\tau} E[V_0(\tau - u)]\mathrm{d}B(u)\mathrm{d}\tau &= \int_{u=0}^{\infty} b(u) \int_{\tau=u}^{\infty} e^{-s\tau} \cdot E[V_0(\tau - u)]\mathrm{d}\tau\mathrm{d}u \\
&= \int_{u=0}^{\infty} e^{-su} b(u) \int_{y=0}^{\infty} e^{-sy} \cdot E[V_0(y)]\mathrm{d}y\mathrm{d}u \\
&= \int_{u=0}^{\infty} e^{-su} b(u)\mathrm{d}u \cdot \int_{y=0}^{\infty} e^{-sy} \cdot E[V_0(y)]\mathrm{d}y \\
&= b^\star(s) \cdot Q^\star(s)
\end{aligned}$$

and $b^\star(s) := \int_{\tau=0}^{\infty} e^{-s\tau}\mathrm{d}B(\tau)$. We now have an expression for the Laplace transform $Q_0^\star(s)$:

$$\begin{aligned}
Q_0^\star(s) &= \frac{1/s + 2\lambda H^\star(s)}{s + \lambda - \lambda b^\star(s)} \\
&= \frac{1}{s^2} \cdot \frac{1 + 2\lambda s H^\star(s)}{1 + \lambda \cdot \frac{1 - b^\star(s)}{s}}
\end{aligned}$$

Also, let $Q_1^\star(s)$ be the Laplace-Stieltjes transform of $E[V_1(\tau)]$:

$$\begin{aligned} E[V_1(\tau)] &= \int_{x=0}^{\infty} \frac{1-B(x)}{E[B]} E[V_1(\tau,x)]\mathrm{d}x \\ &= \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot (2x + E[V_0(\tau-x)])\mathrm{d}x + 2\tau \int_{x=\tau}^{\infty} \frac{1-B(x)}{E[B]}\mathrm{d}x \\ &= \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot E[V_0(\tau-x)]\mathrm{d}x + 2E[min(R,\tau)], \end{aligned}$$

where $R$ is the distribution of the residual life time of the customer present upon arrival, see for instance [5, p. 111].

$$R(t) := \frac{1}{E[B]} \int_{x=0}^{t} [1-B(x)]\mathrm{d}x$$

Now the Laplace transform $Q_1^\star(s)$ is given by

$$Q_1^\star(s) = \frac{1-sB^\star(s)}{sE[B]} \cdot Q_0^\star(s) + 2\widehat{H^\star}(s),$$

where $\widehat{H^\star}(s) := \int_{\tau=0}^{\infty} e^{-s\tau} \cdot E[min(R,\tau)]\mathrm{d}\tau$ and $B^\star(s) := \int_{\tau=0}^{\infty} e^{-s\tau} \cdot \mathrm{d}B(\tau)$. Note that $sB^\star(s) = b^\star(s)$.

$$\begin{aligned} &\int_{\tau=0}^{\infty} e^{-s\tau} \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot E[V_0(\tau-x)]\mathrm{d}x\mathrm{d}\tau = \\ &\int_{x=0}^{\infty} \int_{\tau=x}^{\infty} e^{-s\tau} \cdot \frac{1-B(x)}{E[B]} \cdot E[V_0(\tau-x)]\mathrm{d}\tau\mathrm{d}x = \\ &\int_{x=0}^{\infty} \int_{y=0}^{\infty} e^{-s(y+x)} \cdot \frac{1-B(x)}{E[B]} \cdot E[V_0(y)]\mathrm{d}y\mathrm{d}x = \\ &\int_{x=0}^{\infty} e^{-sx} \cdot \frac{1-B(x)}{E[B]} \int_{y=0}^{\infty} e^{-sy} \cdot E[V_0(y)]\mathrm{d}y\mathrm{d}x = \\ &\int_{x=0}^{\infty} e^{-sx} \cdot \frac{1-B(x)}{E[B]}\mathrm{d}x \int_{y=0}^{\infty} e^{-sy} \cdot E[V_0(y)]\mathrm{d}y = \frac{1-sB^\star(s)}{sE[B]} \cdot Q_0^\star(s) \end{aligned}$$

Also,

$$\begin{aligned} \widehat{H^\star}(s) &= \int_{x=0}^{\infty} \frac{1-B(x)}{E[B]} \cdot \frac{1}{s^2}[1-e^{-sx}]\mathrm{d}x \\ &= \frac{1}{s^2} \int_{0}^{\infty} \frac{1-B(x)}{E[B]}\mathrm{d}x - \frac{1}{s^2} \int_{0}^{\infty} \frac{1-B(x)}{E[B]} \cdot e^{-sx}\mathrm{d}x \\ &= \frac{1}{s^2}(1 - \frac{1-sB^\star(s)}{sE[B]}) \end{aligned}$$

The last equation can be derived from the following:

$$\begin{aligned} &\int_{\tau=0}^{\infty} e^{-s\tau}[\int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot 2x \cdot \mathrm{d}x + \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot 2\tau \cdot \mathrm{d}x]\mathrm{d}\tau = \\ &\int_{\tau=0}^{\infty} e^{-s\tau} \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot 2x \cdot \mathrm{d}x\mathrm{d}\tau + \int_{\tau=0}^{\infty} e^{-s\tau} \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot 2\tau \cdot \mathrm{d}x\mathrm{d}\tau \end{aligned}$$

$$\begin{aligned}\int_{\tau=0}^{\infty} e^{-s\tau} \int_{x=0}^{\tau} \frac{1-B(x)}{E[B]} \cdot 2x \cdot \mathrm{d}x\mathrm{d}\tau &= \int_{x=0}^{\infty} \frac{1-B(x)}{E[B]} \cdot 2x \int_{\tau=x}^{\infty} e^{-s\tau} \cdot \mathrm{d}\tau\mathrm{d}x \\ &= \int_{x=0}^{\infty} \frac{1-B(x)}{E[B]} \cdot 2x \cdot \frac{1}{s} e^{-sx}\mathrm{d}x\end{aligned}$$

$$\begin{aligned}\int_{\tau=0}^{\infty} e^{-s\tau} \int_{x=\tau}^{\infty} \frac{1-B(x)}{E[B]} \cdot 2\tau \cdot \mathrm{d}x\mathrm{d}\tau &= \int_{x=0}^{\infty} \frac{1-B(x)}{E[B]} \int_{\tau=0}^{x} e^{-s\tau} \cdot 2\tau \cdot \mathrm{d}\tau\mathrm{d}x \\ &= \int_{x=0}^{\infty} \frac{1-B(x)}{E[B]} \cdot (\frac{2}{s^2} - e^{-sx}(\frac{2x}{s} + \frac{2}{s^2}))\mathrm{d}x\end{aligned}$$

To acquire the Laplace transform ($Q^\star(s)$) of $E[V(\tau)]$ we need to average over the possible states of the system where an arriving customer is accepted:

$$\begin{aligned}Q^\star(s) &= \frac{p_0}{p_0+p_1} \cdot Q_0^\star + \frac{p_1}{p_0+p_1} \cdot Q_1^\star \\ &= \frac{1}{1+\rho} \cdot \frac{1/s + 2\lambda H^\star(s)}{s+\lambda-\lambda b^\star(s)} + \frac{\rho}{1+\rho} \cdot (\frac{1-sB^\star(s)}{sE[B]} \cdot Q_0^\star(s) + 2\widehat{H^\star}(s)) \\ &= \frac{1}{1+\rho} \cdot (\frac{1/s + 2\lambda H^\star(s)}{s+\lambda-\lambda b^\star(s)} \cdot (1 + \rho \cdot \frac{1-sB^\star(s)}{sE[B]}) + \rho \cdot 2\widehat{H^\star}(s)) \\ &= \frac{1}{1+\rho} \cdot (\frac{1/s + 2\lambda H^\star(s)}{s+\lambda-\lambda b^\star(s)} \cdot (1 + \frac{\lambda - \lambda(b^\star(s) + B(0))}{s}) + \rho \cdot 2\widehat{H^\star}(s)) \\ &= \frac{1}{1+\rho} \cdot (\frac{1/s + 2\lambda H^\star(s)}{s+\lambda-\lambda b^\star(s)} \cdot (\frac{s+\lambda-\lambda b^\star(s)}{s}) + \rho \cdot 2\widehat{H^\star}(s)) \\ &= \frac{1}{1+\rho} \cdot (\frac{1/s + 2\lambda H^\star(s)}{s} + \rho \cdot 2\widehat{H^\star}(s)) \\ &= \frac{1}{1+\rho} \cdot (\frac{1/s + 2\lambda \cdot \frac{1}{s^2}(1-sB^\star(s))}{s} + \rho \cdot 2 \cdot \frac{1}{s^2}(1 - \frac{1-sB^\star(s)}{sE[B]})) \\ &= \frac{1}{1+\rho} \cdot (\frac{2\rho+1}{s^2} + 2\lambda \cdot \frac{1-sB^\star(s)}{s^3} - \frac{2\rho}{E[B]} \cdot \frac{1-sB^\star(s)}{s^3}) \\ &= \frac{1+2\rho}{1+\rho} \cdot \frac{1}{s^2},\end{aligned}$$

where $p_j := \frac{1}{1+\rho+\rho^2} \cdot \rho^j$ and $\rho := \lambda \cdot E[B] < 1$.
If we perform the inverse Laplace transformation we get:

$$E[V(\tau)] = \frac{1+2\rho}{1+\rho} \cdot \tau$$

Now if we take the limit of $\rho \to 0$ we see that $E[V(\tau)] = \tau$ and if we take the limit of $\rho \to \infty$, then $E[V(\tau)] = 2\tau$, which are both what we intuitively expect.

## 3.5 Symmetric queues and sojourn times

In this section we will show that using the properties of a symmetric queue it is possible to find an expression for the mean conditional sojourn time in the PS queue with admission control. The theory of symmetric queues is discussed

in [10] and [16]. The reader should be warned that again the indices of the parameters are dropped. The parameters in this section refer to those of the data process.

First we will show that the PS queue without admission control is a symmetric queue. After that we will also show that, with slight adjustments, the PS queue with admission control is also symmetric.

Consider a queue which customers are ordered, with the queue containing customers in positions $1, 2, \ldots, n$, where $n$ is the total number of customers in the queue. Customers are assumed to be one of $K$ classes and the service requirement distribution of customers depends on their class. A queue is called symmetric (see [10]) if it operates in the following manner:

- The service requirement of a customer is a random variable whose distribution may depend upon the class of the customer.
- A total service effort is supplied at rate $\phi(n)$.
- A proportion $\gamma(l, n)$ of this effort is directed to the customer in position $l$; when this customer leaves the queue, customers in positions $l+1, l+2, \ldots, n$ move to position $l, l+1, \ldots, n-1$ respectively.
- When a customer arrives at the queue, he moves into position $l$ with probability $\gamma(l, n+1)$; customers previously in positions $l, l+1, \ldots, n$ move to positions $l+1, l+2, \ldots, n+1$ respectively.

For the PS queue this will result in the following:

- $\phi(n) = 1$, $n = 1, 2, \ldots,$
- $\gamma(l, n) = \frac{1}{n}$, $l = 1, 2, \ldots, n$ and $n = 1, 2, \ldots,$

We will also give the $\phi(n)$ and $\gamma(l, n)$ for the Last-Come First-Served (LCFS) discipline:

- $\phi(n) = 1$, $n = 1, 2, \ldots,$
- $\gamma(l, n) = 0$, $l = 1, 2, \ldots, n-1$ and $n = 1, 2, \ldots,$
- $\gamma(l, n) = 1$, $l = n$ and $n = 1, 2, \ldots,$

We now have the building blocks to show that the PS queue with admission control is also symmetric. If customers arrive when the queue has $C$ customers, where $C$ is the maximal queue length, the queue will stop serving in a PS manner. Instead the newly arrived customers will be served in a LCFS manner until the queue length has again dropped to $C$ customers. Now set the service speed of the LCFS discipline to $M$, with $M$ very large. The customers that arrive in a *full system* will be served at such a speed that they will leave the queue almost immediately. The service to the customers already present in the queue will hardly be interrupted and the queue length will not be larger than the admission threshold $C$ for any significant amount of time, if $M$ is large. We can denote this by:

- $\phi(n) = 1$, $n = 1, 2, \ldots, C$
- $\phi(n) = M$, $n = C+1, C+2, \ldots,$
- $\gamma(l, n) = \frac{1}{n}$, $l = 1, 2, \ldots, n$ and $n = 1, 2, \ldots, C$
- $\gamma(l, n+1) = 0$, $l = 1, 2, \ldots, n$ and $n = C+1, C+2, \ldots,$
- $\gamma(n+1, n+1) = 1$, $n = C+1, C+2, \ldots,$

This results in the observation that the PS queue with admission control is also symmetric.

For symmetric queues one can easily obtain the probability $P\{N = n\}$ that there are $n$ customers in the system, see [16]. Let the total arrival rate of work over all classes be

$$\rho = \sum_{k=1}^{K} \rho_k, \tag{3.21}$$

and define

$$\varphi_k = \frac{\rho_k}{\rho} \tag{3.22}$$

The value of $\varphi_k$ is the probability that a given position in the symmetric queue is a class-$k$ customer. From [16] formula (10.68) the probability that there are $n$ customers in the system is given by

$$P\{N = n\} = b \frac{\rho^n}{\phi_1 \ldots \phi_n} \sum_{n_1 + \ldots + n_K = n} \left( \begin{array}{c} n \\ n_1, \ldots, n_K \end{array} \right) \varphi_1^{n_1} \ldots \varphi_K^{n_K}, \tag{3.23}$$

with $b$ denoting an appropriate normalization constant. In the PS queue with admission control, which is described above, this formula simplifies to

$$P\{N = n\} = \begin{cases} b\rho^n & n = 0, 1, \ldots, C \\ b\frac{\rho^n}{M^{n-C}} & n = C+1, C+2, \ldots, \end{cases} \tag{3.24}$$

with $b = [\sum_{n=0}^{C} \rho^n + \sum_{n=C+1}^{\infty} \frac{\rho^n}{M^{n-C}}]^{-1}$, the normalization constant.

If one knows the expected value of the number of customers in the system $E[N] = \sum_{n=0}^{\infty} nP\{N = n\}$, one can find an expression for the expected value of the sojourn time. All terms above $C$ in the summation used to compute $E[N]$ will approach 0, because of the large $M$. Using this, the expression for $E[N]$ can be simplified to summing over all states up to $C$, $n = 0, 1, \ldots, C$. This leads to

$$E[N] = \frac{\sum_{n=0}^{C} n\rho^n}{\sum_{n=0}^{C} \rho^n} = \frac{C\rho^{C+2} - (C+1)\rho^{C+1} + \rho}{(1-\rho)(1-\rho^{C+1})} \tag{3.25}$$

For a class-$k$ customer one knows using Little's Law

$$E[N_k] = \lambda_k E[V_k] \tag{3.26}$$

and Nelson (see [16] and equation (3.22)) states:

$$E[N_k] = \frac{\rho_k}{\rho} E[N] \tag{3.27}$$

Combining these results gives:

$$E[V_k] = \frac{\rho_k}{\rho} \frac{1}{\lambda_k} E[N] = \frac{\beta_k}{\rho} E[N] = \frac{C\rho^{C+1} - (C+1)\rho^C + 1}{(1-\rho)(1-\rho^{C+1})} \beta_k \tag{3.28}$$

Because we want to know the sojourn time for an accepted customer we need to scale the arrival rate $\lambda_k$. The arrival rate of accepted customers is $\lambda_k^a = \lambda_k \cdot P\{N \leq C-1\}$.

$$P\{N \leq C-1\} = \sum_{n=0}^{C-1} b\rho^n = \frac{\sum_{n=0}^{C-1} \rho^n}{\sum_{n=0}^{C} \rho^n} = \frac{1-\rho^C}{1-\rho^{C+1}} \tag{3.29}$$

So the sojourn time of accepted class-$k$ customers in the PS queue is

$$E[V_k^a] = \frac{\rho_k}{\lambda_k^a} \frac{E[N]}{\rho} = \frac{\sum_{n=0}^{C} n\rho^{n-1}}{\sum_{n=0}^{C-1} \rho^n} \cdot \beta_k = \frac{C\rho^{C+1} - (C+1)\rho^C + 1}{(1-\rho)(1-\rho^C)} \cdot \beta_k \tag{3.30}$$

This result is independent of the service requirement distribution. The parameter $\beta_k$ is the mean amount of service a class-$k$ customer requires. One could replace this by a general value $\tau$ to obtain the mean conditional sojourn time of an arbitrary customer.
An interesting limit is the limit of $\rho$ going to infinity and of $\rho$ going to zero.

$$\lim_{\rho\to\infty} E[V(\tau)] = \lim_{\rho\to\infty} \frac{\sum_{n=0}^{C} n\rho^{n-1}}{\sum_{n=0}^{C-1} \rho^n} \cdot \tau = C \cdot \tau \tag{3.31}$$

$$\lim_{\rho\to 0} E[V(\tau)] = \frac{C\rho^{C+1} - (C+1)\rho^C + 1}{(1-\rho)(1-\rho^C)} \cdot \tau = \tau \tag{3.32}$$

These results show that when the system is almost always fully occupied the capacity is divided equally between all customers. Each customer will get $\frac{1}{C}$ of the capacity. If a customer would get all capacity the sojourn time would be equal to $\tau$. With $\frac{1}{C}$ of the capacity it will take $C$ times the service time $\tau$ to complete the transfer.

# Chapter 4

# Numerical results and conclusions

In section 4.1 we will present a number of graphs showing the results of various parts of this thesis. We will show the effects of adding extra timeslots to the system and give the allowed voice load and data arrival rate ($\lambda_d$) for a fixed number of timeslots. Also a brief description of the main results will be given in section 4.2.

## 4.1 Numerical results

We look at a GPRS communication system with voice and data traffic. We model this system through several 1-dimensional queueing systems for data traffic linked through a quasi-stationary regime. The number of 1-dimensional queues for data traffic is equal to the number of states of the voice process. Voice traffic is modeled by a $M/G/c_v/c_v$ queueing system and data traffic by a $M/G/1/c_d - PS$ queueing system. We want to know the maximal voice load given a certain blocking probability as a QoS constraint. We also want to know the steady-state distribution of the voice process given a certain voice load. The steady-state probabilities of the voice process are used in the quasi-stationary regime. For data traffic we want to compute the mean conditional sojourn time and the throughput given a certain available capacity. The capacity for data customers is the left over capacity by voice customers of the total capacity. This capacity is known if the number of voice customers is known, see equation (2.4).

### 4.1.1 Maximal voice load under QoS constraints

In this subsection the maximal voice load given a blocking probability of 0.01 is presented. The results are obtained through solving the voice load from setting equation (2.1), with $n = c_v$ timeslots, equal to the blocking probability. This is done using the solve function of Mathematica to solve $\rho_v$ from the following

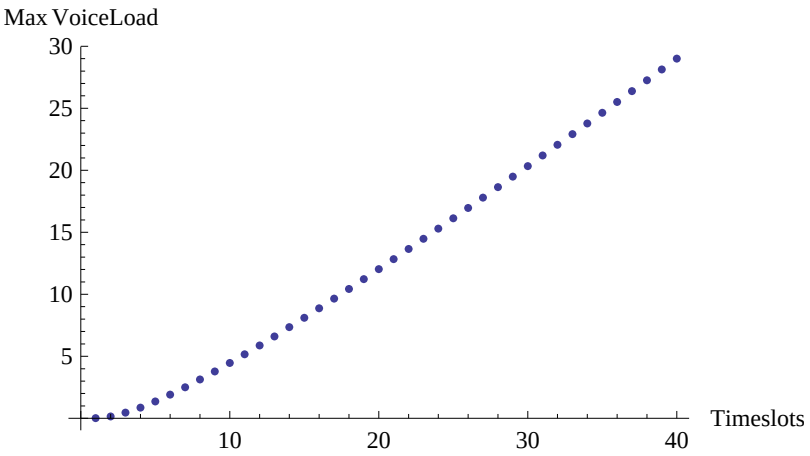


Figure 4.1: Maximal voice load vs. number of timeslots

equation

$$p_v(c_v) = \frac{\frac{\rho_v^{c_v}}{c_v!}}{\sum_{n=0}^{c_v} \frac{\rho_v^n}{n!}} = 0.01,$$

which is the same as setting equation (2.2) equal to 0.01.

We started with a system with $c_v = 1$ timeslot and computed the maximal voice load and continued to do this for systems with $c_v = 2, 3, \ldots, 40$ timeslots available. In the figure 4.1 one can see the results.

Figure 4.1 shows a nearly linear growth, except in the first part. Adding extra timeslots will result in a higher maximal voice load under the same QoS constraints.

### 4.1.2 Differential equation approach

In this subsection we will present the numerical results for the mean conditional sojourn time of data customers in the $M/M/1/c_d - PS$ queue. We will give the functions of the mean conditional sojourn time $E[V_i(\tau)]$ of an arriving customer seeing $i$ data customers in the system on arrival. These results come from the theory described in section 3.2 and we solve equation (3.19) using the software Mathematica. The following parameter settings are used in the system of differential equations: $c_d = 4$, $\lambda = 2$, $\mu = 4$ and $r = 1$. This represents a system with 4 timeslots available for data traffic. Here $\tau$ is the amount of service a data customer requires. In this model we looked at the mathematical properties. The $\lambda$, $\mu$ and $r$ are given numerical values just so we could solve the differential equations.

$$E[V_0(\tau)] = -0.243704 + 0.000139393e^{-22.8489\tau} - 0.00545455e^{-10.\tau} + 0.249019e^{-3.15114\tau} + 1.73333\tau$$

$$E[V_1(\tau)] = 0.122963 - 0.00145309e^{-22.8489\tau} + 0.0218182e^{-10.\tau} -$$

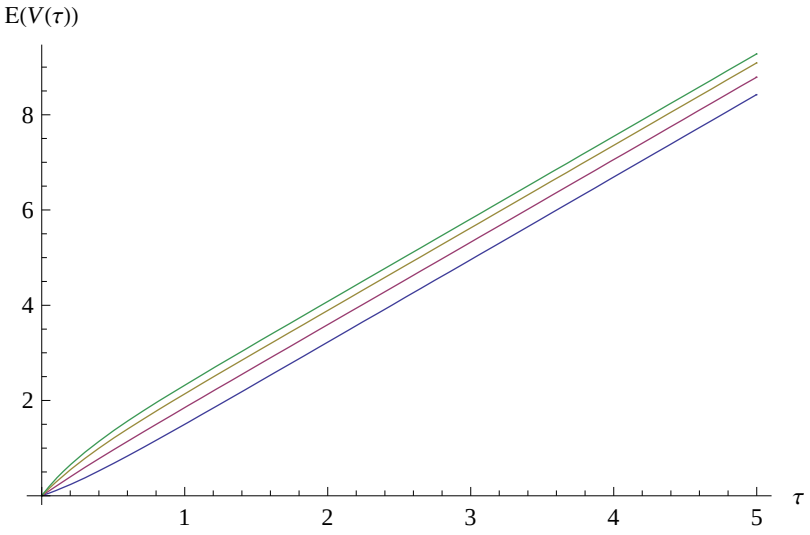


Figure 4.2: Mean conditional sojourn time of an arriving customer vs. service requirement

$$0.143328e^{-3.15114\tau} + 1.73333\tau$$

$$E[V_2(\tau)] = 0.422963 + 0.00525478e^{-22.8489\tau} - 0.00545455e^{-10.\tau} - 0.422763e^{-3.15114\tau} + 1.73333\tau$$

$$E[V_3(\tau)] = 0.611852 - 0.00581235e^{-22.8489\tau} - 0.0327273e^{-10.\tau} - 0.573312e^{-3.15114\tau} + 1.73333\tau$$

In figure 4.2 one can see that all functions are (nearly) linear for large values of $\tau$. There is a small difference between the sojourn time of customers arriving in an empty system and customers who see other customers on arrival. This difference remains constant for very large $\tau$ and is negligible compared to the value of the sojourn time. The differential equation approach is useful when one wants to make statements about the difference in sojourn time of a customer seeing either $n$ or $m$ other customers upon arrival. For small values of $\tau$ and having a lot of timeslots available the difference in sojourn time between entering an empty system or a full system could be significant.

### 4.1.3 Symmetric queue approach

In this subsection we will present numerical results for the mean conditional sojourn time using the symmetric queue approach described in section 3.5. The mean conditional sojourn time is computed using equation (3.30) with $\rho = \frac{\lambda}{cr\mu}$ and $\beta_k = \frac{\tau}{cr}$. By defining $\rho = \frac{\lambda}{cr\mu}$ we make some assumptions on the model. In this case the capacity of the queue is embedded in the term $cr$ in $\frac{\lambda}{cr\mu}$. In this model the admission threshold for data customers is defined by $c$. We

state that there can be a maximum of $c$ voice customers and data customers simultaneously present in the system, which is equal to the number of timeslots in the system. If there is 1 voice customer present in the system there can be a maximum of $c-1$ data customers present. So, if the number of data customers in the system is called $c_d$ and the number of voice customers in the system is called $c_v$ they obey the following equation

$$c = c_v + c_d.$$

The capacity for data customers obeys the following equation

$$C_d = C_{total} - C_v = c \cdot r - c_v \cdot r$$

In this section we adjust the admission threshold AND the capacity for data customers according to the number of voice customers present. This happens because $c$ is used inside $\rho$ and as the $C$ from equation (3.30). To get numerical results we produced some functions and procedures in mathematical software. We will briefly discuss each of the used functions.

The first function "**EVtRate**" is equation (3.30) from section 3.5. The function "EVtRate" computes the mean conditional sojourn time (sec) of an accepted customer in a $M/G/1/c-PS$ queue.

$$\mathbf{EVtRate}[\tau, c, \lambda, \mu, r] := \frac{\tau}{cr} \frac{c(\frac{\lambda}{cr\mu})^{c+1} - (c+1)(\frac{\lambda}{cr\mu})^c + 1}{(1-(\frac{\lambda}{cr\mu}))(1-(\frac{\lambda}{cr\mu})^c)} \tag{4.1}$$

Variables "**EVtRate**":

$\tau$ = total amount of work for the accepted customer (kbit)

$c$ = number of timeslots available for data customers (-)

$\lambda$ = arrival rate of data customers (1/sec)

$\frac{1}{\mu}$ = mean amount of work for data customers (kbit)

$r$ = processing rate per timeslot (kbit/sec)

Next the function "**VP**", which is equation (2.1), computes the equilibrium distribution of the voice queue ($M/G/c/c$).

$$\mathbf{VP}[i, \rho, c] = \frac{\frac{\rho^i}{i!}}{\sum_{n=0}^{c} \frac{\rho^n}{n!}} \tag{4.2}$$

Variables "**VP**":

$i$ number of customers in the voice queue (-)

$\rho$ load of the voice process (-)

$c$ number of timeslots available for voice customers (-)

The function "**EVqs**" computes the quasi-stationary sojourn time (sec) of a data customer. The theory around the quasi-stationary regime can be found in section 3.1.

$$\mathbf{EVqs}[\tau, c, \lambda, \mu, r, \rho] = \sum_{i=0}^{c-1} \mathbf{VP}[i, \rho, c] \cdot \mathbf{EVtRate}[\tau, c-i, \lambda, \mu, r] \tag{4.3}$$

Variables "**EVqs**":

$\tau$ total amount of work for the accepted customer (kbit)

$c$ number of timeslots available for data and/or voice customers (-)

$\lambda$ arrival rate of data customers (1/sec)

$\frac{1}{\mu}$ mean amount of work for data customers (kbit)

$r$ processing rate per timeslot (kbit/sec)

$\rho$ load of the voice process (-)

The function "**THtRate**" computes the throughput (kbit/sec) of data customers in the $M/G/1/c-PS$ queue. The throughput is computed through means of the mean conditional sojourn time of a data customer in the $M/G/1/c-PS$

$$\mathbf{THtRate}[\tau, c, \lambda, \mu, r] = \frac{\tau}{\mathbf{EVtRate}[\tau, c, \lambda, \mu, r]} \tag{4.4}$$

$\tau$ total amount of work for the accepted customer (kbit)

$c$ number of timeslots available for data customers (-)

$\lambda$ arrival rate of data customers (1/sec)

$\frac{1}{\mu}$ mean amount of work for data customers (kbit)

$r$ processing rate per timeslot (kbit/sec)

The function "**THqs**" computes the quasi-stationary throughput (kbit/sec) of data customers. Again, for the theory of the quasi-stationary regime see section 3.1.

$$\mathbf{THqs} = \sum_{i=0}^{c-1} \mathbf{VP}[i, \rho, c] \cdot THtRate[\tau, c-i, \lambda, \mu, r] \tag{4.5}$$

$\tau$ total amount of work for the accepted customer (kbit)

$c$ number of timeslots available for data customers (-)

$\lambda$ arrival rate of data customers (1/sec)

$\frac{1}{\mu}$ mean amount of work for data customers (kbit)

$r$ processing rate per timeslot (kbit/sec)

$\rho$ load of the voice process (-)

The procedure "**FindMaxLambda**" computes the data arrival rate ($\lambda_d$) for which the quasi-stationary data throughput is equal to a certain target value. It uses bi-section to find this data arrival rate ($\lambda_d$).

$$
\begin{aligned}
\textbf{FindMaxLambda}\quad & [\tau, c, \mu, r, \rho, THTarget, output] = \\
& (ll = 0; \qquad lr = 1000; \qquad test = 100; \\
& WHILE[\qquad ABS[test] > 0.0001, \\
& \qquad\qquad l = (ll + lr)/2; \\
& \qquad\qquad TH = \textbf{THqs}[\tau, c, l, \mu, r, \rho]; \\
& \qquad\qquad test = THTarget - TH; \\
& \qquad\qquad IF[test < 0, ll = l, lr = l]; \\
& ]; output = l)
\end{aligned}
$$

$\tau$ total amount of work for the accepted customer (kbit)

$c$ number of timeslots available for data customers (-)

$\frac{1}{\mu}$ mean amount of work for data customers (kbit)

$r$ processing rate per timeslot (kbit/sec)

$\rho$ load of the voice process (-)

$THTarget$ value of the target throughput (kbit/sec)

*output* dummy variable

To perform the dimensioning of the communication system one first needs to compute the voice load ($\rho_v$) that makes the blocking probability equal to the QoS constraint. This voice load will be the maximal voice load that is allowed under the QoS constraint. Now for each $\rho_v$ between zero and the maximum voice load one needs to find the maximal allowed arrival rate ($\lambda_d$) of the data customers. This can be done using **FindMaxLambda**. Now we will present results for the dimensioning of the system. For this we need to give numerical values to the parameters $\tau, c, \mu, r$ and $THtarget$.
Set $\tau = 100$ kbit, $\frac{1}{\mu} = 12$ kbit, $r = 10$ kbit/sec , $THtarget = 20$ kbit/sec and $c = 4$ (figure 4.4), $c = 5$ (figure 4.3) respectively. In these figures one can see the maximal arrival rate of data customers $\lambda_d$ given a certain voice load $\rho_v$. One can see that when the system gets more capacity, by adding timeslots, the maximal allowed $\lambda_d$ is higher. Both figures show that the allowed $\lambda_d$ decreases linearly when $\rho_v$ is increased. If one computes the $\lambda_d$ for $\rho_v = 0$ and $\rho_v = maxvoiceload$ the area underneath a straight line between these point represents the allowed points in the dimensioning. By computing only two points and drawing a straight line between them, the dimensioning could be done with less computing power.

We will take a closer look at the capacity of both systems. In the case of 4 timeslots with a rate of 10 kbits/sec the total capacity of the system is 40

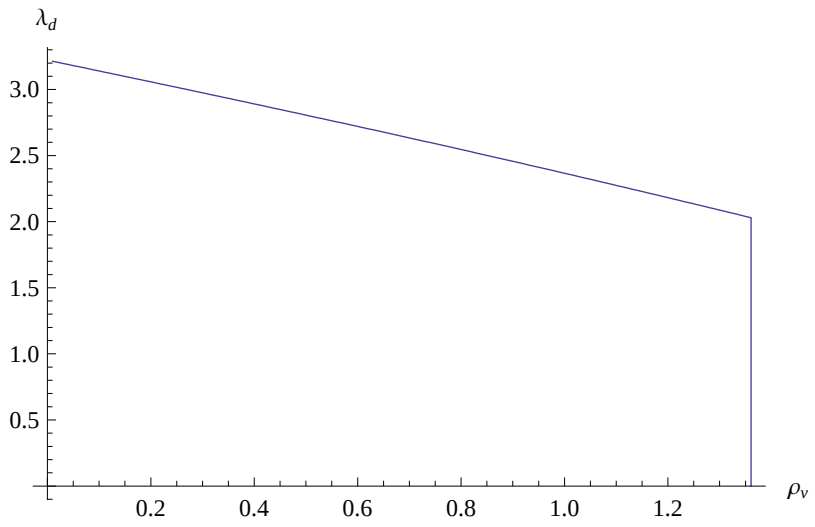


Figure 4.3: Dimensioning, $c = 5$,

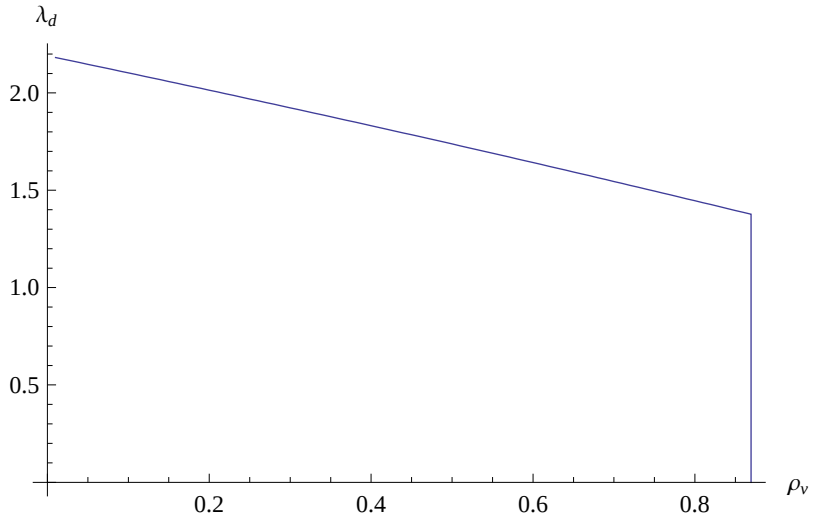


Figure 4.4: Dimensioning, $c = 4$

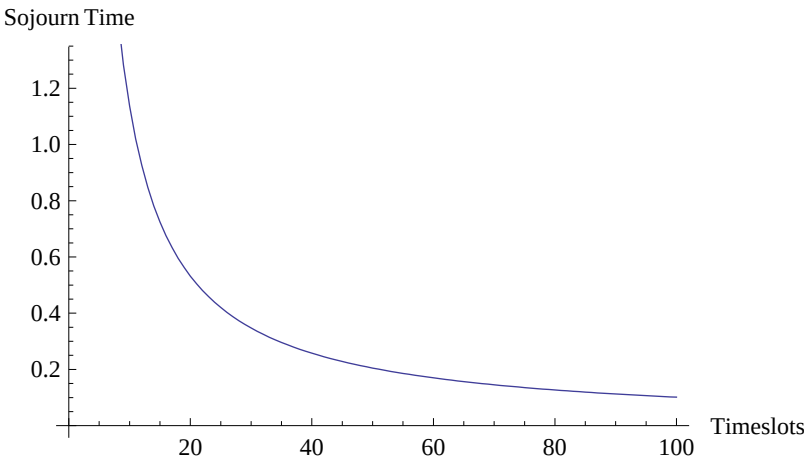


Figure 4.5: Sojourn time vs. number of timeslots

kbits/sec. The arrival rate of work in this system, when the voice load is 0.01, is $2.18 \cdot 12 = 26.12$ kbit/sec on average. At maximal voice load (= 0.8694) the arrival rate of work is $1.3768 \cdot 12 = 16.5216$ kbit/sec on average. For a system with 5 timeslots these numbers are $3.2139 \cdot 12 = 38.5668$ kbit/sec when the voice load is 0.01 and $2.0297 \cdot 12 = 24.3564$ kbit/sec at maximal voice load (= 1.3608). One can see that adding extra timeslots while the QoS constraints remain constant, increases the dimensioning area significantly.

### 4.1.4 Sojourn time as function of the number of timeslots

In this subsection we will present results for the mean conditional sojourn time as function of the number of timeslots. The mean conditional sojourn time is computed through equation (3.30) with $\rho = \frac{\lambda}{cr\mu}$ and $\beta_k = \frac{\tau}{cr}$.

In figure 4.5 the following parameters are used: $\tau = 100$ kbit ,$\lambda = 1$ 1/sec,$\frac{1}{\mu} = 12$ kbit and $r = 10$ kbit/sec. One can see from figure 4.5 that adding timeslots to the system lowers the sojourn time. For the first few extra timeslots the drop in sojourn time is significant. If there are more than around 20 timeslots available the drop in sojourn time per extra timeslot gets smaller and smaller.

### 4.1.5 Data customer arrival rate as function of the number of timeslots

We are interested in the data customer arrival rate ($\lambda_d$) as function of the number of timeslots when the throughput-target ($THtarget$) is fixed and the voice load is either the maximal allowed voice load under QoS constraints or is fixed at 0.000001. From these graphs one could see the effect of adding extra timeslots the system on the allowed arrival rate of data customers. When using the maximal voice load and a fixed throughput-target one could see the effect of adding extra timeslots on the allowed arrival rate of data customers under QoS constraints in voice traffic peak. When using 0.000001 as voice load one

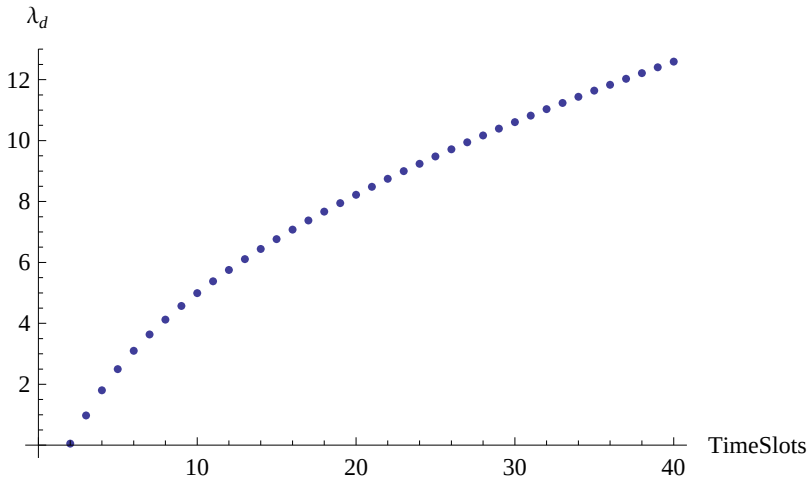


Figure 4.6: Data customer arrival rate at maximal voice load vs. number of timeslots

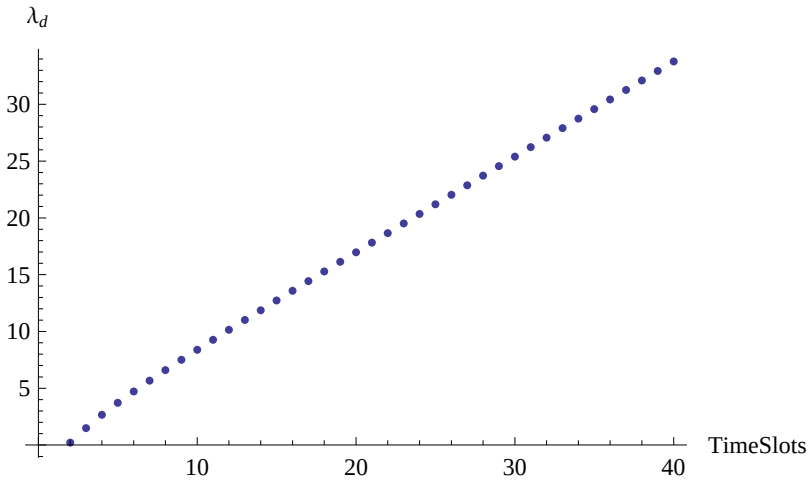


Figure 4.7: Data customer arrival rate low voice load vs. number of timeslots

could see the data customer arrival rate under QoS constraints when there are nearly no voice customers requesting service. First we will present the graph (figure 4.6) of $\lambda_d$ as function of the number of timeslots with the maximal voice load, $\tau = 100$, $\mu = \frac{1}{12}$, $r = 10$ and $THtarget = 18$. Second the graph (figure 4.7)of $\lambda_d$ as function of the number of timeslots with the voice load fixed at 0.000001, $\tau = 100$, $\mu = \frac{1}{12}$, $r = 10$ and $THtarget = 18$ is presented.

In the first graph (figure 4.6) , where the voice load is maximal, the line is curved. In the second graph (figure 4.7) this is not the case. The data customer arrival rate $\lambda_d$ is almost linear when the number of timeslots is larger than 10. Also in the case of nearly no voice traffic the data customer arrival rate $\lambda_d$ increases much faster. One could say that mainly voice traffic is responsible for the curvature in the figures. Still this curvature effect becomes less significant when the number of timeslots is large. As shown in section 4.1.1, the maximal

voice load increases nearly linearly when the number of timeslots is large.

## 4.2 Conclusions

In this section we will give the conclusions of this thesis. The main result of this thesis is the use of the quasi-stationary regime in combination with the Processor-Sharing (PS) queue for dimensioning of the system. In dimensioning one needs to find the maximal allowed voice load and data customer arrival rate $\lambda_d$ under QoS constraints. The effects of the random environment imposed by the voice process are handled through use of the quasi-stationary regime. By using the quasi-stationary regime it is possible to model the integrated system in terms of several 1-dimensional queueing systems. For voice traffic we use the classical $M/G/c_v/c_v$ model and for data traffic we used a $M/G/1/c_d - PS$ model.

We showed in chapter 2 that in isolation the voice traffic is easily dimensioned through the classical Erlang loss model. Also for the PS useful results are found for the mean conditional sojourn time for the data process in isolation. In chapter 3 we extend the PS model to handle admission control and combine the voice and data process with the quasi-stationary regime. Under the quasi-stationary regime the mean conditional sojourn time of the data customers again gives a linear result when conditioning on the file size. We showed that through systems of differential equations one can find solutions for the mean conditional sojourn time. This approach can also be used to find solutions for $E[V_i(\tau)]$, the mean conditional sojourn time for a data customer seeing $i$ voice customers in the system upon arrival. With this one can make statements about the difference in sojourn time when entering a nearly full or empty system. Unfortunately this approach still uses a lot of computation power and is hard to solve analytically for bigger systems. For this reason we turned to an approach using the characteristics of symmetric queues. With this approach we were able to find a solution for $E[V(\tau)]$ for a certain number of available timeslots. The symmetric queue approach also gives solutions which are easily computed.

# Bibliography


[1] J.S. Belrose, *On the birth of wireless telephony*. Available at `http://www.telecommunications.ca/WirelessTelephony_-_2.pdf`.

[2] J.L. van den Berg, O.J. Boxma, *Sojourn times in feedback and processor sharing queues*. Teletraffic Science for new Cost-Effective Systems, Network and Services, ITC 12, ed. M. Bonatti (North-Holland Publ. Cy., Amsterdam).

[3] T. Bonald, S. Borst, N. Hedge, A. Proutière, *Wireless Data Performance in Multi-Cell Scenarios*. REPORT PNA-E0302 Probability, Networks and Algorithms, November 14 ,2003 (CWI Amsterdam)

[4] J.W. Cohen, *The Multiple Phase Service Network with Generalized Processor Sharing*. Acta Informatica 12 (1979), 245-284.

[5] J.W. Cohen, *The Single Server Queue (2nd ed.)*. North-Holland, Amsterdam (1982).

[6] F. Delcoigne, A. Proutière and G. Régnié, *Modelling integration of streaming and data traffic*. Proceedings of the ITC specialist seminar on Internet traffic engineering and traffic management (2002), Würzburg, Germany.

[7] A.K. Erlang, *The theory of probabilities and telephone conversations*. Nyt tidsskrift for matematik B, vol. 20 (1909), 33-39.

[8] A.K. Erlang, *Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges*. P. O. Electrical Engineers Journal (1917), 10:189.

[9] M.T.S. Jonckheere, *Personal website CWI of Matthieu Jonckheere*, Available at `http://homepages.cwi.nl/ jonckhee/RR.html`. Retrieved at July 15, 2009.

[10] F.P. Kelly, *Reversibility and Stochastic Networks*. Wiley, Chichester (1979).

[11] D.G. Kendall, *Some problems in the theory of queues*. J. Roy. Statist. Soc Ser. B 13 (1951), 151-185.

[12] L. Kleinrock, *Queueing Systems, Vol. II: Computer Applications*. Wiley, New York (1976).

[13] R. Litjens, *Capacity Allocation in Wireless Communication Networks, -Models and Analyses-*. Febodruk, Enschede, The Netherlands (2003) PhD Thesis University of Twente.

[14] J.D.C. Little, *A Proof for the Queuing Formula:* $L = \lambda W$. Operations Research 9, No. 3, (1961) 383-387.

[15] R. Lloyd-Evans, *QoS in integrated 3G networks*. Artech House, Boston - London (2002).

[16] R. Nelson, *Probability, stochastic processes, and queueing theory*. Springer-Verlag (1995).

[17] R. Núñez Queija, *Processor-Sharing Models for Integrated-Services Networks*. Ph.D. thesis, Technische Universiteit Eindhoven, The Netherlands, 2000.

[18] T.J. Ott, *The sojourn-time distribution in the M/G/1 queue with processor-sharing*. J. Appl. Prob. 21, (1984) 360-378.

[19] B.A. Sevast'yanov, *An ergodic theorem for Markov processes and its application to telephone systems with refusals*. Theory of Probability and its Applications, Vol. 2 (1957), 104-112.

[20] R.W. Wolff, *Poisson arrivals see time averages*. Operations Research 30 (1982), 223-231.

[21] Wikipedia Eigenvalue, eigenvector and eigenspace. (2009, 10 July). In Wikipedia, the free encyclopedia. Available at `http://en.wikipedia.org/wiki/Eigenvalue,_eigenvector_and_eigenspace`. Retrieved July 10, 2009.

[22] S.F. Yashkov, *A derivation of the response time distribution for an M/G/1 processor sharing queue*. Problems Contr. & Info. Theory 12, (1983) 133-148.

[23] S.F. Yashkov, *A note on asymptotic estimates of the sojourn time variance in the* $M/G/1$ *queue with processor-sharing*. Syst. Anal. Model. Simul. 3, (1986) 267-269.

[24] S.F. Yashkov, *The moments of the sojourn time in the M/G/1 processor sharing system*. Institute for Information Transmission Problems, Moscow, Russia. Received June 15, 2006.