CAPACITY ALLOCATION IN WIRELESS COMMUNICATION NETWORKS

- MODELS AND ANALYSES -

ISBN 90-9017132-0

Cover design by Remco Litjens, Maria Fleuren and Coen de Vos. The cover exemplifies the modelling component of wireless network performance analysis with a gradual transition of the realistic 'best server areas' to their typical hexagonal modelling equivalent.

Printed by Febodruk, Enschede, The Netherlands

CAPACITY ALLOCATION IN WIRELESS COMMUNICATION NETWORKS

- MODELS AND ANALYSES -

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 12 september 2003 om 16:45 uur

door

REMCO LITJENS

geboren op 19 oktober 1970
te Arcen

Dit proefschift is goedgekeurd door:

Promotor	prof. dr. W.H.M. Zijm
Assistent-promotor	dr. R.J. Boucherie

Aan Mirjam

CONTENTS

PREFACE

THIS monograph presents the achievements of almost five strenuous years of part-time doctoral research carried out at KPN Research (now: TNO Telecom). At this point, I wish to thankfully acknowledge the invaluable support that has been offered by various people and in different forms.

Soon after I joined KPN Research, my colleague Eric Smeitink pointed out an on-going STW project on 'Stochastic network analysis for the design of self-optimising cellular mobile communications systems' of Richard Boucherie of the Universiteit van Amsterdam. The project was readily assessed to offer a promising opportunity for my doctoral adventure. Well before I was quite aware of my commitment, an appreciated and unconventionally rich budget was formally approved by KPN Mobile and KPN Research, thanks to the determined efforts of Eric Smeitink and my manager Marc Hofstra.

The collaboration with Richard Boucherie that led to the completion of this monograph proved demanding yet tremendously enriching. I greatly benefited from his knowledge in stochastics and his exceptional meticulousness. While Richard's characteristic no-nonsense and critical association could be relentless at times, it effectively trained me to thoroughly evaluate and defend each aspect of our joint work. One memorable incidence was the time when a post-it note on my telephone said: *'Bewijs Lemma 5, blz. 12, 8^e '='-teken is fout'* (fortunately, the equality turned out to be correct). Our inherently distinct objectives ensured that the cooperation never became dull: whereas my principal aspiration is to somehow resolve relevant problems in wireless networking, Richard is primarily driven by the mathematical challenges of stochastic modelling and analyses, with wireless networking merely as an appealing application area. As I sincerely believe that all these experiences and differences made me grow as an applied researcher, I am grateful for all aspects of our pleasant and fruitful cooperation.

Henk Zijm is thanked for acting as my promotor and for his constructive comments. I am further grateful to all other members of my promotion committee for their thorough examination of the lengthy manuscript. Sindo Núñez Queija offered an indispensable basis for the work presented in Chapters 2 through 6 with his valuable monograph [173] and suggestions. The semi-annual comments and directions given by the user committee of the above-mentioned STW project are also appreciated.

At KPN Research/TNO Telecom, I am particularly indebted to Hans van den Berg, head of the Center Of Excellence QOS, who is an enduring source of inspiration and guidance, and admirably committed to exploit his strong theoretical expertise with an ever practical perspective. Our agreeable cooperation has contributed to various chapters in this monograph. Furthermore, I thankfully acknowledge Kevin Kennedy, Maria Fleuren and Frank Roijers for their contributions to Chapters 3 and 9. All other 'wireless' (ex-)colleagues are appreciated for the valuable discussions and fine collaborations in the past five years, which enhanced my general understanding of wireless communication networks: Anton, Bart, Bi, Ellen, Elles, Eric, Erik, Katherina, Ljupčo, Neill, Onno, Renske, Stefan, Stefan, Suzana, Willem, while still others are co-responsible for the exceptional working atmosphere. The 'external' colleagues at KPN Mobile, Swisscom CT, Telia Research, the NTT DoCoMo R&D Center, Ericsson/WMC, the COST 279 action and the Universiteit Twente are thanked for all constructive interactions.

Although all the (ex-)room mates at the various offices are naturally admired for enduring my presence, I owe a particular word of gratitude to Rob Kooij, for his general enthusiasm, a persistent eagerness to resolve any gap in mathematical proofs and a healthy sense of perspective. The cover design is a joint conception with Coen de Vos and Maria Fleuren, while I thank Gerard de Groot of KPN Mobile for providing a screendump of the GSM planning tool. KPN Research/TNO Telecom is thanked for the stimulating environment and excellent facilities. For the time spent elsewhere, I appreciate the hospitality and facilities granted to me by the Universiteit van Amsterdam, the Centrum voor Wiskunde en Informatica and the NTT DoCoMo R&D Center, Japan.

Some of the rare moments that I was not working, were well-spent on the soccer field, volleyball or tennis court. The structural clustering of all these activities on Mondays made for an exhausting yet time-efficient start of the week. I thank my fellow sports(wo)men for all the sweat and joy.

At a personal level, I greatly appreciated the interest expressed and distractions offered by family, in-laws and friends, whom I was often regretfully forced to neglect. I am particularly thankful for my ever-supportive family. *'Ik bof maar met jullie'*, as my brother Mike rightfully put it in his monograph [147]. My father's interest was furthermore conveniently exploited when he kindly agreed to proofread the 'Samenvatting'. I dedicate this monograph to my wonderful wife Mirjam. Mirjam, if it wasn't for your unconditional love and support, and the optimal environment you offered, this monograph would contain merely empty pages. You deserve our upcoming travels more than anyone. Thanks!

Remco Litjens
Voorschoten, August 2003

CHAPTER 1

INTRODUCTION

WITHIN a mere two decades, the wireless networking revolution has unfolded from scratch to the current scenario where the global mobile community counts over a billion users. Although the principal service is still speech telephony, it is generally anticipated that within another decade, data transfers and multimedia communications will become the dominant wireless service mode. With the gradual emergence of the third-generation networks and the development of a range of attractive services, such as multimedia messaging, video conferencing, location-based and transaction services, the wireless communications market continues to expand.

Whereas the primary objective in the development of wireless networks is to devise a system that provides a high spectrum efficiency with a large degree of flexibility, the main task of a wireless network operator is to plan and operate its network most cost-efficiently, while enabling service providers to deliver adequate yet affordable service quality to the end-customer. The modelling and performance analysis of wireless networks provides essential insights for the system development phase, as well as concrete input for the derivation of planning guidelines and the optimisation of control parameters. Such modelling and analysis is undeniably imperative, as wireless networks become inherently more complex due to technology advancements, the offered telecommunication services become more diverse in their characteristics and quality requirements, and the drive for cost-efficiency seems ever more relevant.

In this monograph, we concentrate on capacity allocation in cellular and Wireless Local Area Networks from a network operator's perspective. An insightful reference model is presented for the suite of capacity allocation mechanisms that apply at different time scales. With an objective to comprehend and, ultimately, to optimise the joint impact of these mechanisms on the network operations and service provisioning, we offer contributions in three distinct areas. Anchored in a solid base of technological knowledge, we capture the essential aspects of the considered wireless networks into *tractable models*, develop tailor-made *performance evaluation* methods, and exploit these analyses in *numerical experiments*, in order to expose and quantify the inherent trade-off between network capacity and service quality.

The outline of this chapter is as follows. In Section 1.1 the evolution of wireless communication networks is described with a particular focus on cellular and Wireless Local Area Networks in Europe. Subsequently, Section 1.2 clarifies some basic aspects of wireless communications, concentrating on the network technologies and models studied in this monograph. A discussion of the main challenges in the planning and operations of a wireless communication network is given in Section 1.3. Section 1.4 then presents a structured overview of the principal capacity allocation mechanisms that are applied in wireless communication networks in order to manipulate the above-mentioned trade-off. A general statement of contribution and an outline of the applied performance evaluation methods is given in Section 1.5. Section 1.6 ends this introductory chapter with an overview of the investigations contained in the remainder of this monograph.

1.1. EVOLUTION OF WIRELESS COMMUNICATIONS

The evolution of wireless communications began in 1864 when 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 is thus credited with inventing modern radio communications. Shortly thereafter, Aleksandr Stepanovich Popov and Guglielmo Marconi independently demonstrated the electromagnetic transmission and reception of messages in 1895. The world's first patent on wireless telegraphy using Hertzian waves was awarded to Marconi in 1896 but overturned by the US Surpreme Court in 1943 in favour of Nikola Tesla, after 30 years of legal battles.

While the commercially oriented Marconi deemed Morse code telegraphy sufficiently adequate for communication between ships and across oceans, the earliest experiments with wireless telephony were left to Reginald Aubrey Fessenden and still others. Fessenden recognised that continuous wave transmission was required for speech telephony rather than the spark-based signals generated by Marconi. On December 23, 1900, he generated the first-ever intelligible speech successfully broadcast by radio waves, and not withstanding the short distance and the poor quality of the transmission, this date heralded the beginning of wireless telephony.

Over the years, wireless speech telephony was further enhanced, initially using Amplitude Modulation (AM) and later the more robust Frequency Modulation (FM) scheme, developed by the American inventor Edwin Howard Armstrong in 1935, which formed the technological basis for the first analogue cellular networks. For a more elaborate overview of the early history of wireless communications, we refer to e.g. [15, 106, 145, 163]. In line with the focal technologies of this monograph, the remainder of this section concentrates on the specific evolution of cellular and Wireless Local Area Networks.

1.1.1. CELLULAR NETWORKS

In 1979, the Japanese telecommunications operator NTT deployed the world's first public wireless telephony network in Tokyo based on the cellular concept. Two years later, the cellular era reached Europe when Nordic Mobile Telephone (NMT) systems became operational in Scandinavia, which later spread in slightly different versions to several countries in Europe. These first-generation wireless Metropolitan or even Wide Area Networks (MAN/WANs) all applied Armstrong's FM-based *analogue* modulation. At the end of the 1980s, Europe was equipped with several distinct cellular systems that were generally unable to interoperate [138].

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, which was transferred to the European Telecommunications Standards Institute (ETSI) in 1989 [89, 159]. Three years later, after a delay due to the non-availability of approved terminals ('God Send Mobiles'), the official commercial launch of the second-generation GSM (now: Global System for Mobile communications) took place (see Figure 1.1). In order to indicate GSM's *global* success story, at the end of 2002, 467 GSM networks were deployed in 169 countries, serving 787.5 million customers, about 70% of the world's wireless market [94]. The main features of the second-generation *digital* systems, with GSM as principal exponent, are superior speech quality, enhanced spectral efficiency, low terminal, operational and service costs, a high security level, international roaming, support of low-power hand-portable terminals and a variety of new services. GSM is an evolving standard, and the PHASE 1 services of speech telephony, Short Message Service (SMS, small data chunks up to 140 bytes), facsimile and Circuit-Switched Data (CSD) transfer at 9.6 kbits/s are supplemented by additional services in subsequent phases [171, 192].

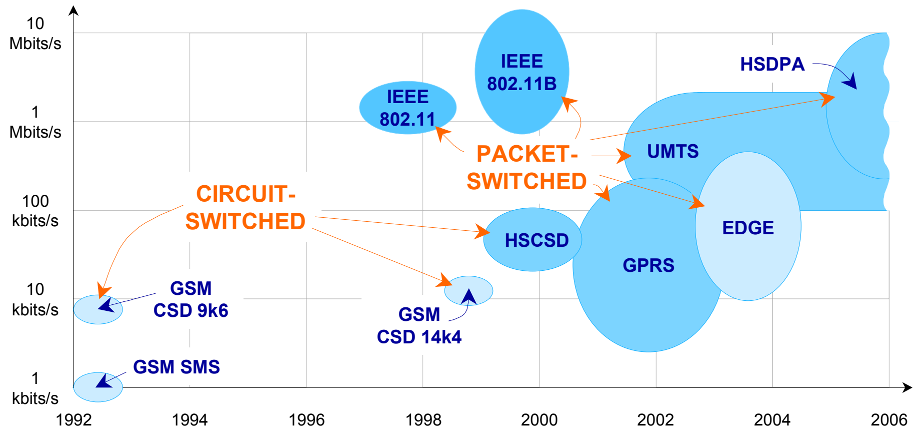


Figure 1.1 The recent evolution of cellular and Wireless Local Area Networks. The technologies that are indicated with a darker ellipse are considered in this monograph.

Among the upgrades of GSM's data transfer capabilities that are specified in PHASE 2+ in order to support the growing market for mobile data services, a new data channel coding scheme with reduced overhead and hence poorer error protection has been standardised to offer a 14.4 kbits/s CSD information bit rate. Secondly, the HSCSD (High-Speed Circuit-Switched Data) service enables data calls to be assigned multiple traffic channels in parallel in order to enhance data rates. An advanced HSCSD can technically achieve an aggregate data transfer rate of $8 \times 14.4 = 115.2$ kbits/s by bundling a maximum of 8 traffic channels [69, 71]. Aside from e.g. large file transfers, HSCSD is also suitable for delay-sensitive applications, given the inherent reliability and small delay variations that are due to its circuit-switched character. The deployment of HSCSD in an existing GSM radio access network merely requires a software update and is thus relatively cheap. Currently, 38 GSM operators in 27 countries worldwide have successfully implemented the HSCSD upgrade, making it available to over 100 million customers [94].

As a third and most significant PHASE 2+ upgrade of the GSM network, GPRS (General Packet Radio Service) introduces packet-switching into the cellular network [37, 39]. The packet-switched character enables the dynamic and flexible sharing of multiple traffic channels by multiple (typically bursty) data flows, in order to enhance service quality and establish a high resource efficiency. Another benefit is the possibility of sensible charging on a per volume basis, which makes it affordable to be always on-line and hence facilitates quick access. The maximum data rate that an advanced GPRS terminal can theoretically attain in favourable propagation conditions is equal to $8 \times 21.4 = 171.2$ kbits/s. One drawback of GPRS compared to HSCSD, which builds entirely on the existing infrastructure, is that its deployment requires major soft- and hardware upgrades to the access and core networks, and is thus significantly more costly. Currently, GPRS services are offered by 167 GSM network operators in 59 countries [94].

As a final significant second-generation network upgrade, EDGE (Enhanced Data rates for Global Evolution) 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 [86, 87]. An EDGE-enhanced second-generation network is occasionally suggested as a viable yet much cheaper alternative to an entirely new third-generation network. Currently, EDGE is either deployed or scheduled for future deployment by 29 network operators in 22 countries worldwide [94].

The anticipated market for wireless multimedia communications, the associated need for additional capacity, the technological advances, and a desire for global roaming, have all driven the development of a new universal standard for mobile cellular networks. A principal element of the IMT-2000 (International Mobile Telecommunications-2000) family of 3G standards, the Universal Mobile Telecommunications System (UMTS) is designed in the 3rd-Generation Partnership Project (3GPP), a joint project of the standardisation bodies from Europe, Japan, Korea, China and the USA. The significant flexibility and efficiency of UMTS are in essence due to its packet-oriented character and an entirely new radio interface based on wideband Code Division Multiple Access (CDMA) technology [218]. While CDMA also underlied a number of second-generation cellular networks that are deployed e.g. in Japan, Korea and the USA, the radio access technology is now sufficiently mature to be applied in the *universal* third-generation standards. In UMTS data rates up to 2 Mbits/s are technically possible [45, 59, 105, 138, 176]. Currently, 114 3G licences have been awarded worldwide [216]. On October 1, 2001, the Japanese network operator NTT DoCoMo commercially launched the world's first 3G network based on wideband CDMA technology. With the initial growth slower than expected, the NTT DoCoMo 3G network counts just over 150.000 subscribers at the time of writing. The European rollout of UMTS networks is currently on-going, initially deploying 'UMTS islands in an ocean of GSM', but gradually expanding towards national coverage.

A number of technological improvements of the initial UMTS system release are standardised under the name High-Speed Downlink Packet Access (HSDPA), which may be regarded, at least to some extent, as the third-generation equivalent of EDGE. Driven by the generally anticipated up/downlink traffic asymmetry where the bulk of the wireless traffic is expected to comprise of data flows from the base station to the mobile terminals, the main objective of HSDPA in UMTS networks is to enable the support of downlink peak rates in the range of $8-10$ Mbits/s for best effort packet data services, by means of a number of enhanced technologies, e.g. higher order modulation, fast link adaptation, fast scheduling, hybrid link layer retransmissions, fast cell selection and Multiple-Input Multiple-Output (MIMO) antenna technology [2, 185].

1.1.2. WIRELESS LANS

Whereas the cellular network evolution has been driven mainly by representatives of the traditional telephony industry, the development of Wireless Local Area Networks (WLANs) is primarily managed by the data communications industry. Since the first products appeared around 1990 after the worldwide release of the unlicensed ISM (Industrial, Scientific, Medical) spectrum, the WLAN market grew on the basis of the appropriateness of the wireless solution to specific applications, regarding its ease of installation, flexibility, and support for mobility. This growth brought forward a plethora of several proprietary WLAN standards. The harmonising efforts of the Institute of Electrical and Electronics Engineers (IEEE) led to the release of the international IEEE 802.11 WLAN standard in 1997, which has been developed as part of the family of 802.X LAN standards including the 802.3 Ethernet and 802.5 Token ring [110].

This first IEEE 802.11 standard comprised of three distinct physical layer technologies (including one based on infrared transmission), which frustrated both vendors and customers desiring genuine interoperability. The Wireless Ethernet Compatibility Alliance (WECA) was formed in 1997 to compel vendors to jointly ensure interoperability, which led to the ratification of the next generation IEEE 802.11B (or WIFI: WIreless FIdelity) standard in 1999 [111]. The IEEE 802.11B standard supports data rates of 1, 2, 5.5 and 11 Mbits/s using a singular spread spectrum physical layer technology (as also applied in UMTS; see below), and created a tidal wave of momentum from both traditional and new WLAN vendors [189] (see also Figure 1.1).

Aside from the IEEE 802.11B standard, the IEEE 802.11A standard was ratified in the same year for deployment in a higher frequency band, achieving up to 54 Mbits/s. In Europe the IEEE 802.11A products are currently still unlawful as the spectrum allocations are not fully established yet, and the standardisation of the required transmission power control and dynamic channel selection schemes is still ongoing in IEEE's task group TG-H. Further enhancements of the IEEE WLAN standards are developed within a number of other task groups. For instance, enhanced service quality control features are being specified in task group TG-E in support of service integration and differentiation, an inter-access point protocol is devised in task group TG-F, in order to better support terminals as they migrate between the service ranges of different access points, while task group TG-G is examining high-speed extensions to IEEE 802.11B.

1.1.3. INTEGRATION OF CELLULAR NETWORKS AND WLANS

Whereas wide area cellular networks are traditionally speech-oriented and WLANs are typically installed as (extensions to wireline) office data networks, the evolution of services and the desire for seamless roaming underlies the significant current interest in the integration of both network types. In an integrated scenario, cellular networks are deployed in order to provide wide area coverage, while the WLANs access technology is used to cost-effectively boost capacity in local hot spots such as airports or conference centres. As Figure 1.2 illustrates, such a scenario is in line with the relative coverage ranges of cellular and Wireless Local Area Networks. The desired degree of seamlessness in the network integration depends on various aspects, e.g. the expected degree of terminal mobility between access networks, and whether both access networks are controlled by the same operator.

1.2. BASIC ASPECTS OF WIRELESS COMMUNICATIONS

In this section we highlight some aspects that are fundamental to the development and operations of a wireless communication network. Undeniably, the primary resource for wireless communications is *spectrum*. A *multiple access* scheme transforms the spectral resources into physical channels that can be assigned to specific

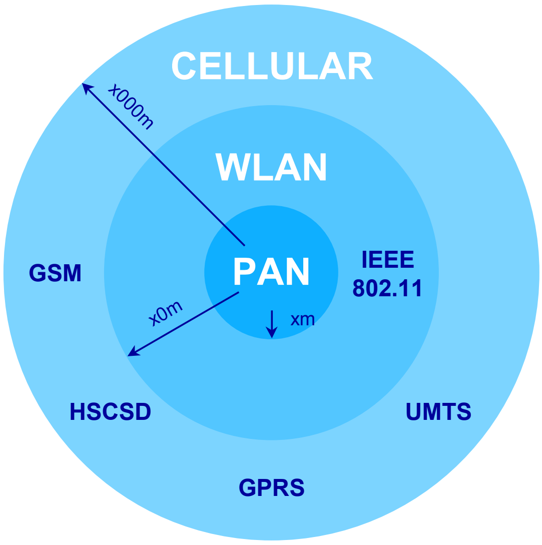


Figure 1.2 Coverage ranges of a base station in a wide area cellular network (e.g. GSM/HSCSD/GPRS, UMTS: high power, long range), an access point in a Wireless Local Area Network (e.g. IEEE 802.11/b: medium power, medium range) and a transmitting node in a Personal Area Network (e.g. Bluetooth: low power, short range).

telecommunication sessions. Such a physical connection comprises of (typically) electromagnetic waves that *propagate* across the wireless medium from the transmitter to the intended receiver. In the following, these elemental aspects are considered in more detail, with a focus on the wireless network technologies and models that are dealt with in this monograph.

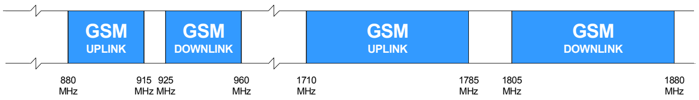


Figure 1.3 Spectrum allocations for the GSM/HSCSD/GPRS band.

1.2.1. SPECTRUM

Spectral resources are rather intangible yet fundamental to wireless communications. According to [192], ‘spectrum space is probably the most limited and precious resource available in the industrialised world’. Although in principle each nation is in charge of spectrum allocation, nowadays it is often coordinated at e.g. the European or even global level, in order to allow cost-efficient production of communication equipment as well as international roaming.

The spectrum assigned for cellular networks such as GSM and UMTS is *licensed* to the network operators for a specific period of time. As shown in Figures 1.3 and 1.4, the assigned spectrum for both GSM (900 and 1800 MHz bands; including the HSCSD/GPRS/EDGE enhancements) and UMTS networks (2 GHz band) consists of a *downlink* and an *uplink* segment [45, 105, 171, 192], used for the communications from a base station to a wireless terminal and the reverse link, respectively. This up/downlink separation in the frequency domain is referred to as *frequency division duplexing* (FDD). As a tradition, the technically more challenging *higher* frequency band is applied in that direction where *(i)* a larger transmission power budget is available to overcome the innately more severe signal attenuation, and *(ii)* higher costs can be made in the development of the associated more complex transmitters. While for satellite networks these are the earth stations (∼ uplink), in land-mobile cellular networks these are the base stations (∼ downlink). Observe that part of the UMTS spectrum is unpaired, which is used in a *time division duplexing* (TDD) mode, where up/downlink transmissions are separated in the time domain.

The ISM band in the 2.4 GHz frequency range (see Figure 1.4) is *unlicensed* and can thus be used without specific governmental permission, although some regulations are imposed regarding e.g. transmission powers. The ISM band is used for e.g. microwave ovens, baby phones, Bluetooth short range transmissions and for IEEE 802.11 WLANs. While ISM band transmissions are inherently prone to external interference

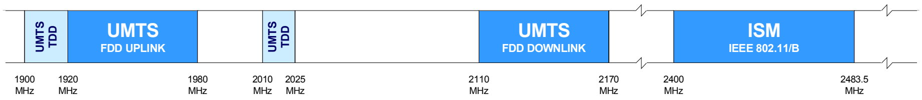


Figure 1.4 Spectrum allocations for the UMTS and ISM bands.

sources due to the band's unlicensed status, the primary advantage is that it can be used free of charge. As is typical for technologies that are intended for relatively short range transmissions, the duplex operation of the IEEE 802.11 standard is established in the time domain [189].

1.2.2. MULTIPLE ACCESS

In first-generation analogue cellular networks, the assigned spectrum was sliced into narrowband (generally non-overlapping) frequency pairs of typically 25 or 30 kHz carrier spacing, with a single conversation occupying one duplex channel. This technique is called *frequency division multiple access* (FDMA). In second-generation networks, with the advent of digital radio communications, it became possible and more efficient to generate wider radio carriers (GSM: 200 kHz) and partition these in the time domain into a number of disjoint periodic trains of time slots (TDMA: *time division multiple access*), where each such physical channel is able to support a single conversation (see Figure 1.5) [171, 192].

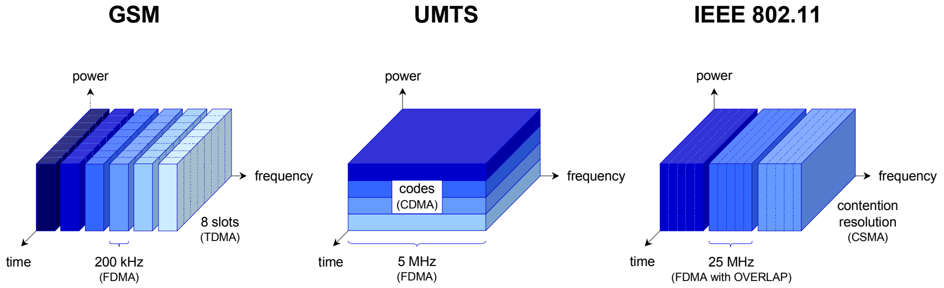


Figure 1.5 The different multiple access schemes standardised for the GSM, UMTS and IEEE 802.11 WLAN technologies.

In any cellular system, a key concept that exploits the benefits from signal strength attenuation, is *frequency reuse* in sufficiently distant cells, also referred to as *space division multiple access* [160]. The spatial allocation of frequency pairs to base stations in a cellular network must conform to an appropriate frequency reuse distance, i.e. as small as possible in order to maximise cellular capacity, yet large enough to prevent harmful interference of concurrent transmissions. A number of advanced techniques, such as discontinuous transmission, slow frequency hopping and reuse partitioning have been developed for second-generation cellular networks in order to enhance radio link qualities with an ultimate objective of increasing cellular capacity by allowing denser frequency reuse.

An entirely different scheme, *code division multiple access* (CDMA) has been selected for the UMTS radio interface standard [105, 138, 218]. In CDMA systems the narrowband information signals are *spread* onto a wideband carrier using a (quasi-) orthogonal channelisation code. These channelisation (or spreading) codes are specifically designed with little or no cross-correlation, in order to minimise the distortion caused by concurrently transmitted signals. Although the primary reason for going through the process of spreading and despreading signals is to enable the CDMA multiple access scheme, an added benefit of the wider bandwidth is the enhanced robustness of signals against frequency-selective fading (see below).

Just as in TDMA-based systems, cellular CDMA-based systems such as UMTS overlay their CDMA scheme on an FDMA plan. As Figure 1.5 indicates, the 5 MHz UMTS carriers are significantly wider than those for GSM. The interference tolerance of spread spectrum transmission enables *universal frequency reuse:* the available carriers can be reused in each cell. As a consequence, signals are concurrently transmitted at the same frequency throughout the network. This resolves the resource inefficiency that is inherent to narrowband FD/TDMA-based networks in the following sense. Whereas in e.g. GSM networks an idle channel in a given cell is generally unavailable to a requested call in another cell, and a temporary inactivity of a speech conversation in between talk spurts implies a waste of capacity, the available capacity of a CDMA-based network is primarily determined by the experienced aggregate interference levels, and hence the capacity variations directly follow the activity fluctuations throughout the network. Another important benefit of the contiguous frequency reuse is that it opens up the possibility of *macro-diversity* or soft handover, in which case a mobile terminal that is typically located near the edge of adjacent cells, can be connected to two or more base stations simultaneously in order to enhance radio link quality at the least favourable locations. Aside from the higher implementational complexity of the scheme, another notable drawback of the CDMA scheme compared to e.g. a TDMA scheme is that the latter multiple access technology provides a higher spectrum efficiency (expressed in capacity per Hz) if only a small area needs to be covered, e.g. in a single cell scenario.

In contrast to the multiple access schemes used in cellular networks, which are typically under centralised control, the *carrier sense multiple access* scheme (CSMA) standardised for the IEEE 802.11 WLAN technology, features distributed control and hence no coordinating entity is required [110, 111, 189]. In the CSMA contention resolution scheme, a terminal is allowed to transmit once it senses that the wireless medium is idle for at least some randomly sampled backoff time, where the random backoff arrangement is used to minimise and resolve conflicts in medium contention. An FDMA-type scheme is specified by the IEEE 802.11 protocol to partition the unlicensed ISM spectral band into thirteen partially overlapping frequencies, among which three non-overlapping frequencies can be identified. While fully overlapping cells (*'basic service areas'* in WLAN terminology) should apply non-overlapping frequencies, a slight spectral overlap is bearable in case of adjacent cells, thus allowing a sparser frequency reuse scheme using five rather than only three frequencies. Different physical layer implementations are possible, including e.g. spread spectrum modulation.

1.2.3. SIGNAL PROPAGATION

An essential aspect of wireless communications is the degradation of radio signals as they propagate from a transmitter to an intended receiver. A general distinction is made between three mutually independent multiplicative propagation phenomena that affect the transmitted signal at different scales (of time and magnitude) [142, 145]. See also Figure 1.6.

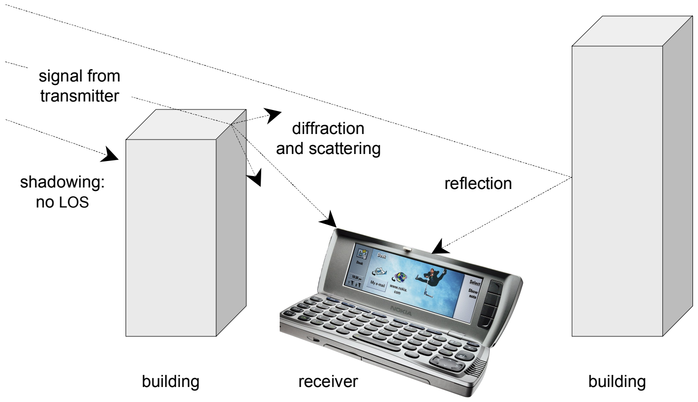


Figure 1.6 Aspects of signal propagation.

At the largest scale, the degree of *attenuation* is predominantly determined by the transmission path length, the antenna heights and the carrier frequency. In generic analyses, the effects of attenuation are usually modelled by assuming an average attenuation which increases with distance according to a power law. *Shadowing* is a medium scale effect which occurs whenever there is an obstruction in the direct path (LOS: Line Of Sight) from the transmitter to the receiver, and is generally caused by obstacles such as buildings or hills. As shadowed areas tend to be large, resulting in a relatively slow rate of change of the signal strength, shadowing is also referred to as ‘slow fading’, or ‘lognormal fading’, which is due to the experimental observation [65] that the local mean power is distributed lognormally around the attenuation-based area mean power. A received signal generally consists of a combination of attenuated, reflected, scattered and diffracted replicas of the original signal, whose different transmission paths and consequently distinct phases, can cause constructive or destructive interference at the receiver. As such *multipath fading* effects vary over very short (typically half-wavelength) distances, the term ‘fast fading’ is often used. A commonly applied model for the statistical time-varying nature of the received instantaneous signal power in cellular networks is the Rayleigh distribution [145].

1.3. CHALLENGES IN CAPACITY ALLOCATION

The principal purpose of capacity allocation in wireless communication networks is to efficiently provide sufficient capacity in order to meet certain customer satisfaction targets where and when needed. Since wireless capacity is intrinsically expensive due to the scarcity of spectrum and the high investment costs, it is of paramount importance for a network operator to allocate the available resources in an efficient and hence cost-effective manner, where ‘resources’ refers to e.g. assigned spectrum, a pool of traffic channels or the available transmission power, depending on the specific context and time scale of the considered capacity allocation problem.

From the perspective of a wireless *network operator*, the relevant network performance indicators are e.g. effective resource utilisation, the amount of carried traffic or plainly revenues. The *customer* satisfaction constraints are expressed in terms of the desired Grade (GOS) and Quality Of Service (QOS). Here the GOS refers to the likelihood that a requested call is served, and is expressed in terms of performance measures such as the call blocking or dropping probability, whereas the QOS refers to the degree of contentment that is associated with a served call, measured by e.g. the ‘outage probability’ that a signal fails to meet its carrier-to-interference ratio (C/I) target, the experienced Bit (or BLock) Error Rate (BER/BLER), transfer time or throughput.

Generally speaking, the key challenges of capacity allocation are imposed by the fact that a network operator is generally unaware of the specific characteristics of the offered traffic in both a temporal and spatial sense. In particular in wireless communication networks, the traffic characteristics consist of a considerable range of distinct, largely uncontrollable, aspects at multiple time scales that all bear some degree of significance to the allocation of capacity. At the largest time scale, the spatial and temporal traffic load fluctuations of different service types can still be predicted to some extent based on market research and extrapolation of experience. At smaller time scales however, the actual instances and locations of call origination, terminal mobility, call burstiness and the variabilities induced by signal propagation are much less predictable yet very relevant aspects. As will be explained in more detail below, specific capacity allocation mechanisms are devised to deal with these uncertain aspects, e.g. handover control handles terminal mobility, Transmission Power Control (TPC) schemes deal with the effects of signal propagation, and appropriate Call Admission Control (CAC) and scheduling enable QOS provisioning despite the unpredictable traffic characteristics.

Analytical and simulation-based performance evaluation studies are carried out extensively to devise and tune the capacity allocation algorithms, using appropriate *probability distributions* and *stochastic processes* in order to capture the unpredictable nature of the offered traffic. The primary objective of such performance investigations is to determine the pertinent trade-off between quality and capacity, which are closely interrelated. As an example, a reduction of the frequency reuse distance in GSM networks increases the number of frequencies and thus the capacity per cell, at the cost of a degradation of the radio link quality due to the increased interference levels. Analogously, loosening the Call Admission Control conditions in a UMTS network effectively raises capacity yet also diminishes the radio link quality. A thorough qualitative yet primarily quantitative understanding of these capacity/quality trade-offs is essential for the standardisation of new technologies, the implementation by vendors and the network planning and operations by network operators. Generally, network operators are allowed substantial freedom to balance the capacity/quality trade-off in line with their positioning strategy in the wireless communications market.

In conclusion, the great potential diversity of services and the corresponding unpredictable traffic characteristics and QOS requirements, the uncontrollability of e.g. terminal mobility and signal fading, and the complex dependencies between the various mechanisms, make capacity allocation a challenging and particularly arduous task of optimisation under uncertainty. As it is generally not viable to carry out true performance optimisation studies that include all relevant aspects in tremendous detail, the mission is to analyse manageable subproblems in order to generate relevant insights into the principal challenges of capacity allocation.

1.4. CAPACITY ALLOCATION MECHANISMS

In this section we provide a structured overview of the different capacity allocation mechanisms that exist to enable the efficient provisioning of the required system capacity and service quality in wireless communication networks. We thus confine ourselves to the mechanisms that apply to the domain of the radio access network operator. Other mechanisms may operate across subnetwork domains, such as end-to-end flow (congestion) control performed by the Transmission Control Protocol (TCP) in order to limit the amount of data in transit and thus restrain data buffer occupancies. [169, 210]. Among the technologies considered in this monograph, we primarily concentrate on the (upgraded) GSM and UMTS cellular networks, as the capacity allocation mechanisms applicable to the IEEE 802.11 wireless local area network technology are rather limited.

As depicted in the reference model of Figure 1.7, capacity allocation mechanisms can be categorised corresponding to the granularity level of operation and the associated time scale (vertical axis). Along the horizontal axis the mechanisms are linked to the radio network entit(y)(ies) in charge of the operations. The involved network entities are indicated with the nomenclature of the GSM, UMTS and IEEE 802.11 access technologies, where the terminal is referred to as the MS (Mobile Station), UE (User Equipment) or STA (station), the base station is called the BTS (Base Transceiver Station), NodeB or AP (Access Point), and the entity in control of the radio network is the BSC/PCU (Base Station Controller/Packet Control Unit) or RNC (Radio Network Controller), respectively. While the left-most column in the figure specifies the principal traffic, customer and environmental aspects that influence the mechanisms at each time scale, the right-most column indicates the associated class of performance measures.

At the network and cell level, the *network planning* mechanisms set the capacity and thus establish the operational framework for the *traffic management* mechanisms that operate at the call and burst/packet level. The objective of traffic management is to exploit these resources most efficiently while providing adequate service quality.

Conversely, the traffic management mechanisms determine the amount of resources services of various types consume, which is an essential input for network planning. In the remainder of this section we elaborate on the various capacity allocation mechanisms. It is noted and illustrated that although each of these mechanisms is designed to carry out a well-defined task, all of them generate and/or utilise resources and hence strong interdependencies exist, as is generically indicated above. As general sources on network planning and traffic (or radio resource) management, we refer to [105, 138, 188, 201, 231], while some mechanism-specific references are included in the text below.

1.4.1. NETWORK PLANNING

Executed by an operator's radio network planning department and typically partially automated, network planning sets the capacity at the network and cell levels.

NETWORK LEVEL

The most 'coarse-grained' capacity allocation mechanisms are deployed at the network level. At this level, the primary mechanisms that determine the network's capacity and coverage are summarised below [55, 66, 102, 114, 142, 146, 166].

SITE PLANNING: The selection of locations to install new BTSs/NodeBs/APs.

CELL SECTORISATION: At selected sites multiple directional antennas can be installed, each covering a partition of the 360° range. In light of the relatively high costs associated with the acquisition, licenses and maintenance of sites in a cellular network, cell sectorisation is the cheapest form of cellular densification and thus a preferred instrument to increase network capacity.

FREQUENCY PLANNING: A frequency plan in GSM networks assigns to each BTS one or more frequency pairs, in accordance with the reuse distance constraints imposed by the carrier-to-interference ratio requirements. As indicated above, in CDMA-based UMTS networks universal frequency reuse is possible and in fact most spectrum efficient (e.g. [105, 138, 218]). In IEEE 802.11-based WLANs the assignment of the partially overlapping frequencies to different APs must conform to the potential overlap of the basic service areas [189].

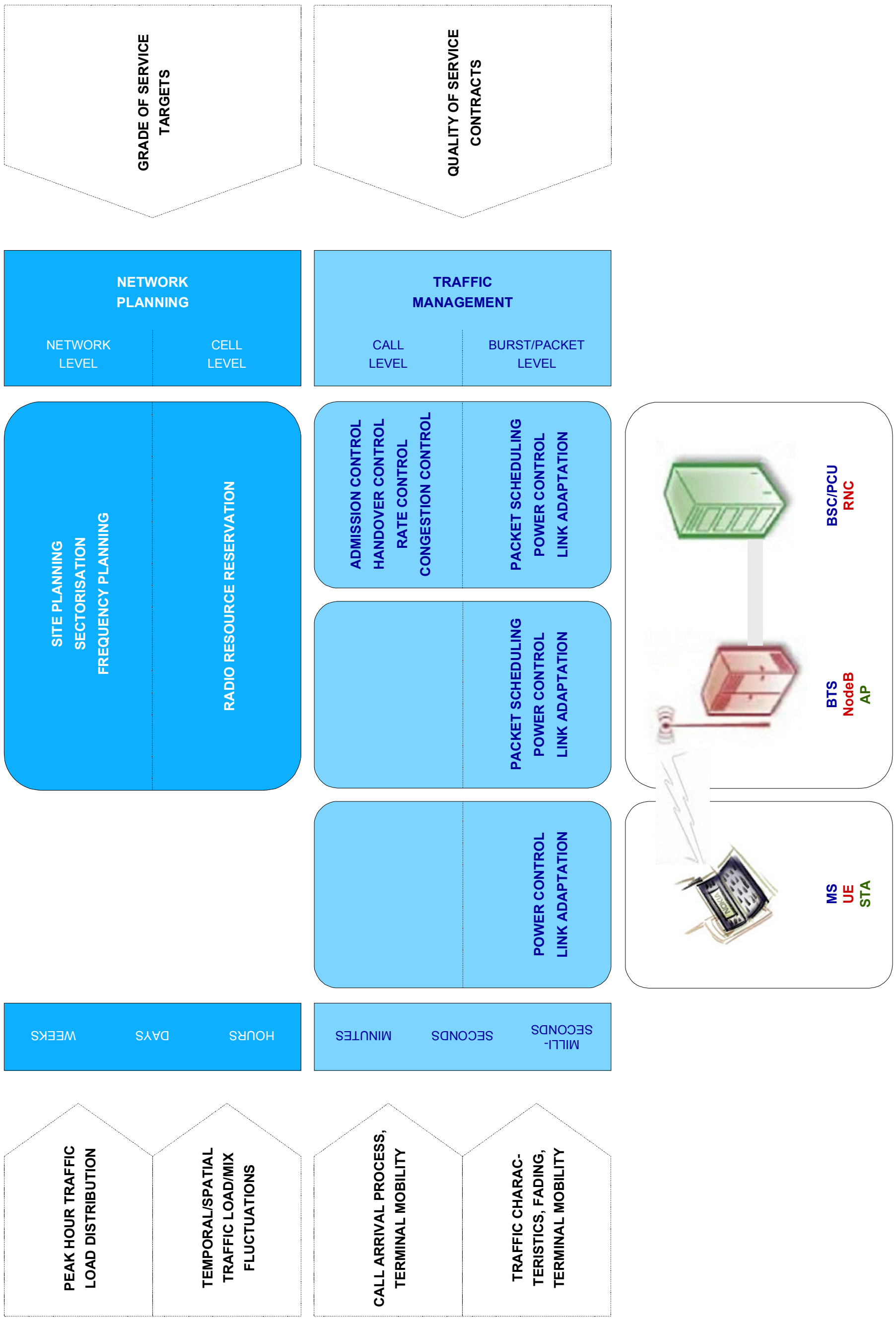


Figure 1.7 Capacity allocation in wireless communication networks: a layered overview of the principal mechanisms and the corresponding time scales.

The mechanisms at the network level are based on the estimated average geographical teletraffic load distribution at peak hours, which in turn follows from market predictions, traffic load measurements and the calculated system capacity. This traffic load distribution must be appropriately matched by the deployed network capacity, in compliance with the operator's policy regarding the GOS targets. Changes at the network level typically occur at a time scale of weeks or even days.

CELL LEVEL

At the cell level, a Radio Resource Reservation policy affects the manner in which the assigned capacity is shared by different call types.

RADIO RESOURCE RESERVATION: In the capacity assignment, RRR policies can be deployed to establish a preference of handover over fresh call requests, 'gold' over 'silver customers' or between different services [23, 136, 144, 217]. As an example of the latter type, an RRR policy in an integrated speech/data network may reserve a fraction of the system capacity strictly for data transfer ($\sim$ trunk reservation), while the remaining capacity is shared with preemptive priority for (delay-sensitive) speech traffic.

The deployed RRR policy and/or its parameters should be in line with the customer- or call-specific GOS targets, and may be dynamically altered on a time scale of e.g. hours or days in response to tempo-spatial traffic load and traffic mix fluctuations.

1.4.2. TRAFFIC MANAGEMENT

Operational in real-time, the key challenge of traffic management is to provide acceptable service quality in the most resource efficient manner.

CALL LEVEL

At the call (or flow) level, we distinguish between call admission, handover, rate and congestion control.

CALL ADMISSION CONTROL: The objective of CAC is to admit or reject fresh or handover call requests, based on the actual capacity, the carried traffic load, the constraints imposed by the RRR policy, and the desired QOS of the newly requested and on-going calls. The CAC mechanism is *proactive*, in the sense that it aims to prevent undesired QOS degradation, and fulfils a key role in the trade-off between capacity (GOS) and service quality (QOS). While in GSM/GPRS networks the design of CAC schemes is comparatively simple, in a CDMA-based UMTS network, it is intrinsically difficult to predict and measure the effect of a newly admitted call on the calls already present and vice versa, given the hard-to-anticipate influence of uncontrollable fading variations and terminal mobility on the experienced QOS. Moreover, CDMA-based networks are characterised by soft capacity, i.e. no 'natural' blocking exists such as in a GSM network where a cell has a fixed number of traffic channels [11, 79, 115, 133]. Although no CAC mechanism is specified in the current IEEE 802.11 standards [110, 111], it is among the enhanced QOS control features specified in task group TG-E [112].

HANDOVER CONTROL: In order to provide uninterrupted access to the inherently untethered wireless terminals, the handover control scheme specifies a set of rules that trigger a request to change a call's serving BTS/NodeB(s), which is submitted to the CAC scheme. Proper handover control ensures that calls are always served by the most suitable BTS/NodeB(s) to ensure resource efficiency, prevent call dropping and provide adequate QOS. It is noted that whereas in GSM networks each terminal is attached to a single BTS only ('break-before-make', *hard handover*), the universal frequency reuse in UMTS networks enables terminals to maintain a so-called 'active set' of multiple NodeBs ('make-before-break', *soft handover*) [24]. While current IEEE 802.11 implementations may have proprietary solutions to support inter-AP handovers, an inter-AP protocol is currently being standardised in task group TG-F [113]. Aside from intra-system handovers also inter-system handovers may be applied in e.g. integrated GSM/UMTS or UMTS/WLAN networks, which is particularly important to ensure coverage in e.g. the initial stages of UMTS deployment or when using WLANs to boost hot spot capacity.

RATE CONTROL: The rate control mechanism exploits the flexibility of elastic (e.g. video or data) services by dynamically adjusting the assigned resources in accordance with variations in prioritised traffic loads [23]. The objective of rate control, which is a form of *adaptive scheduling*, is to protect prioritised traffic when needed, enhance elastic service quality where and when possible, and implicitly optimise resource efficiency.

CONGESTION CONTROL: Primarily in UMTS networks, overload situations may occasionally occur in the sense of intolerable interference levels, typically due to uncontrollable call level dynamics such as terminal mobility. It is the objective of the congestion control mechanism to monitor the network load, detect when rare overload conditions are encountered, evaluate the degree and urgency of the overload conditions, and trigger appropriate counter measures at call and/or burst/packet level in order to get the system quickly but controllably back to a feasible load [46, 202]. Congestion control is of a *reactive* nature, in contrast with the *proactive* CAC mechanism discussed above. Potential counter measures include the rejection of all new call requests (CAC), the denial of power-up commands (TPC, see below), a reduction of data transfer rates (rate control), and as an ultimate resort even controlled call dropping.

Operational at a typical time scale of seconds, call level traffic management mechanisms are primarily influenced by the call arrival process, the induced traffic load variations and terminal mobility.

BURST/PACKET LEVEL

At the burst/packet level a number of traffic management mechanisms operate at the time scale of (milli)seconds.

PACKET SCHEDULING: The packet scheduler handles all non-real-time traffic, i.e. it time- (GSM/GPRS, UMTS) or code-multiplexes (UMTS) the active data flows in order to fulfil such diverse and potentially contradictory aims as resource efficiency optimisation, fairness and QOS differentiation [7, 121]. In an HSDPA-enhanced UMTS network, *adaptive packet scheduling* schemes can be devised that time-multiplex data transmissions based on the flows' rapidly varying fading characteristics at a time scale of milliseconds [35, 156]. In IEEE 802.11 WLANs, the dominant medium access mode is the Distributed Coordination Function, which is based on random access and contention resolution, while the optional Point Coordination Function involves *polling* of data packets by the AP.

TRANSMISSION POWER CONTROL: In GSM networks, the slow and optional TPC mechanisms are able to compensate only for the effects of attenuation and shadowing using updates at a rate of about 2 Hz. In contrast, TPC is of *crucial* importance in UMTS networks, since both capacity and quality are strongly influenced by the experienced interference levels [12]. The joint operations of outer loop (at $10 - 100$ Hz) and fast inner loop TPC (at 1500 Hz) are designed to compensate for attenuation, slow and fast fading fluctuations, in order to minimise the exerted transmission powers such that a call's BLER target is met most efficiently. IEEE 802.11-based WLANs currently feature only *manual* TPC.

LINK ADAPTATION: For data traffic, the link adaptation mechanism dynamically selects the most efficient combination of modulation and channel coding scheme, in response to varying radio link qualities. In GSM/GPRS networks (without the EDGE physical layer enhancements) the modulation scheme is fixed so that only adaptive channel coding can be applied. For UMTS/HSDPA networks it is recommended to serve data flows with a fixed transmission power [184], while the fast link adaptation mechanism substitutes TPC by continuously optimising the modulation and coding scheme in response to the varying channel conditions. In IEEE 802.11-based WLANs no channel coding is applied, leaving only adaptive modulation (*automatic rate fallback* [189]).

While Transmission Power Control and link adaptation are designed to deal with the effects caused by uncontrollable aspects such as terminal mobility and signal fading, the (adaptive) scheduling mechanism responds to both the random traffic characteristics of the active data flows and the fading variations.

1.5. CONTRIBUTION AND APPROACHES

While the overview of the extensive suite of available capacity allocation mechanisms given above provides a mere high-level description of their functionalities, the true challenge of the wireless network operator is to deploy these mechanisms most effectively in a live network. Aside from a thorough technical understanding of the mechanisms' operations, the appropriate use requires quantitative modelling and analysis of a range of identifiable subproblems, which is the primary focus of this monograph.

This section formulates a general statement of the presented contribution in the area of capacity allocation in wireless communication networks, followed by an outline of the applied performance evaluation models and approaches.

1.5.1. GENERAL CONTRIBUTION

The general contribution of this monograph comprises of three distinct aspects that cover all essential phases in performance analysis. *Tractable models* are developed which conveniently abstract from the technological complexities yet still capture the most relevant system and traffic aspects for the performance investigations. Subsequently, novel approaches are developed for a comprehensive *performance analysis* of the constructed models, while existing analytical methods are extended. Lastly, the models and analyses are exploited in a range of purposeful numerical experiments, in order to generate *indispensable insights* in the capacity/quality trade-offs that are inherent to wireless communication networks and of paramount relevance for network operators.

1.5.2. MODELS AND APPROACHES

Although the specifics of the different access technologies considered call for tailor-made evaluation approaches, a few basic stochastic modelling and evaluation techniques can be effectively applied and fruitfully extended to investigate wireless network performance.

LOSS NETWORK MODELS

The most traditional type of stochastic models in the field of telecommunication network performance analysis are the loss network models, which generically feature a pool of resources that are shared by one or multiple services, where the generated calls claim a (service-specific) fixed amount of resources for their autonomously distributed duration [198]. Loss network models are commonly applied to derive GOS performance measures at call level, e.g. fresh call blocking or handover failure probabilities. As a most basic example of a loss network model, the Erlang loss model [67] is still much applied in the planning of speech-oriented communication networks. Another well-known loss model is the multi-service extension of the Erlang loss model, for which a recursive procedure can be derived to determine the loss (blocking) probabilities [124, 194]. Chapter 8 utilises a loss network model to assess the impact of terminal mobility on UMTS network planning. Generally speaking, however, loss network models in their pure form appear to be loosing their applicability to the emerging integrated circuit- and packet-switched networks, although they are expected to remain useful to provide rough first-order insights.

PROCESSOR SHARING MODELS

Originally developed for the analysis of time sharing in computer communication systems, (egalitarian) Processor Sharing (PS) *models* are characterised by a common resource that is fairly shared by a varying number of present jobs, whose sojourn times are determined by the service requirement as well as the experienced amount of service attention [132]. PS models are widely applied to the performance analysis of elastic data transfers in (typically wireline) communication networks, concentrating on QOS performance measures such as throughputs and transfer times (see Chapters 4 and 5). Important generalisations of the basic PS model include the Discriminatory Processor Sharing (DPS) model with multiple service classes [81, 131] (see Chapter 4), and the Generalised Processor Sharing[1] (GPS) model with state-dependent service rates [54], which is applied for data transfer time analyses in GPRS and UMTS networks and WLANs in Chapters 4, 5, 6 and 9, respectively. Chapter 6 considers the performance of multiple interacting PS queues.

INTEGRATED LOSS NETWORK / PROCESSOR SHARING MODELS

As the traditionally dedicated service-specific networks evolve towards inherently more efficient integrated services networks, the evaluation models develop accordingly [152]. In a select number of recent studies an *integration* of loss network and Processor Sharing models is considered to analyse the performance of an integrated system serving speech and elastic (e.g. data) traffic [5, 6, 61, 175]. The complexities of analysing such a hybrid model are due to the integration of services with distinct traffic characteristics and performance metrics. Moreover, the QOS analysis of delay-tolerant elastic traffic is complicated by a fluctuating service rate [173], due to the call level dynamics of typically prioritised speech traffic. Such integrated models are applied in Chapters 2 through 4 for GSM/HSCSD/GPRS-type systems.

[1] Although in this monograph the GPS model is applied in the sense of [54], in the networking community Generalised Processor Sharing generally refers to a scheduling discipline that distributes resources over calls of distinct service classes in proportion to certain class-specific weights, including the provision of minimum bandwidth guarantees to each call regardless of the behavior of the other calls (see e.g. [34, 181, 182]).

OTHER APPROACHES

Another element of the evaluation methods that is common to some of the presented investigations, is the *model decomposition*. For instance, in the UMTS studies of Chapters 6 and 8, a separate analysis of the interference aspects due to terminal location, radio propagation and Transmission Power Control, captures these wireless aspects in a form that can be subsequently applied in a call level analysis of the traffic dynamics. Analogously, in Chapter 9 an isolated packet level throughput evaluation of the WLAN's contention resolution scheme provides relevant input for a flow level PS analysis of the expected data transfer times. Similar decomposition methods are applied in e.g. [79, 137, 139].

Other evaluation techniques that are employed include *Monte Carlo* (Chapters 6, 7) and *dynamic simulations* (used in Chapters 3-6, 8 and 9 in order to complement the analyses).

1.6. OVERVIEW OF THIS MONOGRAPH

In this introductory chapter we have outlined the evolution and described the basic components of wireless communication networks, discussed the principal challenges in capacity allocation and provided a structured overview of the main capacity allocation mechanisms, all with an intentional focus on three distinct network technologies: second-generation GSM networks including the HSCSD and GPRS upgrades, third-generation UMTS networks and IEEE 802.11 wireless LANs. The remainder of this monograph presents a number of performance evaluation studies that are carried out in the specific context of one of these three network types. Although the specifics of the network technology indeed place some non-trivial requirements on the applied models and analyses, important equivalencies exist at the conceptual level. In an outline of the main body of this monograph, the performance analyses are put in perspective using the presented reference model for capacity allocation mechanisms, in order to expose these equivalencies. Figure 1.8 depicts a structural overview of the monograph, where the different chapters are categorised by the considered technology, the focus on single service or integrated services models and whether the applied evaluation method is predominantly analytical or a blend of analysis and simulations. Each chapter includes its own brief introduction, statement of contribution and review of the related literature.

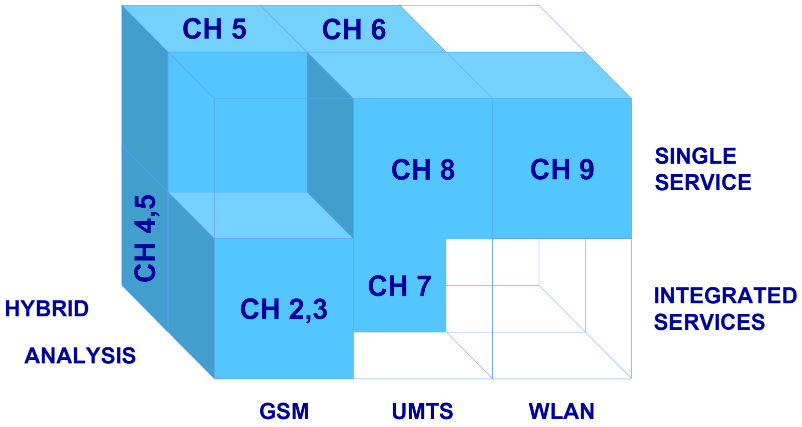


Figure 1.8 Overview of the monograph, structured by the considered technology, the focus on single service or integrated services models and the applied evaluation method.

Chapters 2 through 5 concentrate on extensions of GSM networks. From the perspective of the performance analysis of integrated services GSM/HSCSD/GPRS networks, the relatively distant frequency reuse imposed by the FD/TDMA technology allows us to sensibly concentrate the investigations on a single cell system.

In Chapter 2 we investigate the integration of speech telephony and data transfer in a single GSM/HSCSD cell, where data calls are elastic in the sense that they can handle varying resource assignments. A unified framework for fair *channel sharing* policies is specified, which defines a CAC scheme, an RRR scheme to reserve resources for one or both service types, a rate control scheme to dynamically control the capacity assigned to the data service and a fair PS-based scheduling scheme where the active data calls are assigned equal expected shares of the available capacity. An extensive Markovian analysis is applied to assess and compare the performance of different channel sharing policies. Aside from a number of basic performance measures that can be determined directly from the Markov chain's equilibrium distribution, e.g. the service-specific call blocking probabilities, an analytical expression is derived for the expected sojourn time of an admitted data call, conditional on its size and the system state upon arrival, which may serve as an appreciated feedback information service to the data source. Numerical experiments are included to compare a number of proposed channel sharing policies that fall within the defined unified framework and to obtain insight in the performance effects of the various system and policy parameters.

Chapter 3 extends the model and analysis of Chapter 2 to analyse the performance of a generic GSM/GPRS model including speech, video, high- and low-priority data services. The selected channel sharing policies are based on the outcome of the numerical policy comparison in Chapter 2. While both the data calls and the video calls are of an elastic nature, the principal distinction is that for the video service, a more generous channel assignment and a consequently better throughput enhances the experienced audio/video quality without affecting the call durations, while for data calls the increased throughput implies a reduced sojourn time. As in Chapter 2, a comprehensive (un)conditional Markovian analysis is applied for this extended model, including e.g. the conditional expected video throughput as a function of the call duration and the system state upon arrival.

In view of the observed and commonly acknowledged large degree of variability of e.g. WWW (World Wide Web) page sizes, Chapter 4 relaxes the exponentiality assumption for data call sizes and presents a sensitivity analysis for the (conditional) expected data call sojourn times in a GSM/GPRS network integrating speech and data services. The remarkable observation is presented and analytically supported that in the considered PS model with varying capacity (due to the arrival/departure dynamics of prioritised speech traffic), *the expected data call sojourn time is decreasing in the data call size variability*. This is in contrast with e.g. fixed capacity Processor Sharing or First-In First-Out (FIFO) queueing models, where insensitivity or even the reverse effect holds, respectively. The impact of the data call size variability on the data QOS is further assessed for some of the extended models analysed in Chapters 2 and 3.

As an extension to the throughput analysis for video services presented in Chapter 3, Chapter 5 presents a more general, theoretical analysis of *throughput measures* for elastic calls in PS systems with either fixed or randomly varying service rates. Although the presented analysis is of a generic nature and the obtained insights are certainly more broadly applicable, in our wireless context the fixed capacity scenario models a stand-alone GPRS cell serving elastic (video or data) traffic only, while the case of varying capacity corresponds with a GSM/GPRS cell integrating prioritised speech and elastic calls. A number of distinct throughput measures for the elastic traffic are defined, analysed and compared, e.g. the call-average, time-average and the newly proposed expected instantaneous throughputs. A particular focus is placed on the impact of the elastic call size distribution. The principal conclusion is that across all considered scenarios, the call-average throughput, which is most relevant from the user point of view but typically hard to analyse, appears to be excellently approximated only by the readily derived and rather insensitive expected instantaneous throughput measure.

Chapters 6 through 8 focus on third-generation UMTS networks. Inherent to the designed CDMA-based radio interface are the universal frequency reuse and the resource efficient characteristic that the consumed radio resources depend on the relative terminal locations. The implications for performance analyses are challenging, since unlike for GSM/HSCSD/GPRS evaluations, a sensible UMTS network investigation must consider multiple cells as well as the randomness of terminal location. Note that precisely those system aspects that enhance resource efficiency, significantly complicate the analysis.

Chapter 6 concentrates on data transfer over Downlink Shared CHannels, arguably the most efficient UMTS transport channel type for handling the anticipated volumes of downlink data traffic. In this context, the data traffic load affects the data QOS in two distinct manners. Firstly, a heavier data traffic load implies a greater competition for DSCH resources and thus longer transfer times. Secondly, since each data call served on a DSCH must maintain a so-called Associated Dedicated CHannel for signalling purposes, a heavier data traffic load implies a higher interference level, a higher BLock Error Rate and thus a lower effective aggregate DSCH throughput: *the greater the demand for service, the smaller the aggregate service capacity*. The overall performance analysis is decomposed to segregate the interference aspects from the traffic dynamics. The former stage of interference analysis investigates the effects of terminal location and Transmission Power Control, while the latter stage captures the flow-level dynamics in a multi-dimensional PS model with state-dependent effective service rates that are derived from the results obtained in the former stage. A sequence of gradually more complicated scenarios is evaluated in order to demonstrate the above-mentioned effects of the data traffic load on the experienced QOS.

Chapter 7 focuses on the efficient sharing of resources in a UMTS/HSDPA network integrating speech and data services, and thus pursues analogous objectives as Chapters 2 and 3. However, due to the fundamentally different radio access technology, the involved control operations differ significantly. A common characteristic of the service integration policies investigated for GSM/HSCSD/GPRS networks is the fair sharing of available resources by all data flows (within the same priority class). As will be demonstrated, this *fairness* objective is not trivially extended to a CDMA-based radio interface, due to an inherent difference between fairness in terms of the assigned resources and fairness in terms of the experienced data throughput. The efficiency and fairness objectives are achieved by means of *adaptive scheduling*, which comprises of the balanced employment of rate control and packet scheduling. The main role of rate control is to exploit the characteristic delay tolerance and flexibility of data transfers by downgrading the resources assigned to data transfer in cases of heavy speech traffic, in order to enhance speech performance, and upgrading the assigned resources under light speech traffic to improve data throughputs. At the burst level, the packet scheduler is controlled to operate in line with the selected fairness objective. At the physical layer, the idealised operations of Transmission Power Control and fast link adaptation are assumed for speech and data calls, respectively. The performance of the adaptive schemes is numerically assessed by means of analytical performance optimisation methods in combination with Monte Carlo simulations.

Chapter 8 investigates the *impact of terminal mobility* on UMTS radio network planning, performance and investment costs. This impact stems from two distinct aspects. On the one hand, more severe carrier-to-interference ratio requirements apply in case of higher velocities, due to the combined effects of multipath propagation, Doppler shifts and Transmission Power Control imperfections. As a consequence of the more stringent physical layer requirements, cell sizes must be reduced to maintain the same capacity. On the other hand, a higher degree of terminal mobility requires a greater Radio Resource Reservation regarding call handovers in order to keep the call dropping probability below a prespecified target value. Consequentially, fresh call blocking increases, inducing a further need for denser site planning. An analytical approach is presented which covers capacity allocation mechanisms at all levels, viz. cell size optimisation ($\sim$ site planning), RRR, CAC and TPC, thus demonstrating the relations that exist between the different time scales. The principal strength of the presented model and approach lies therein that it is simple enough to allow a computationally relatively inexpensive performance evaluation and optimisation, yet sufficiently realistic to provide valuable qualitative insight for network planning purposes, as it captures the UMTS network characteristics that are essential to our objectives, i.e. terminal mobility and inter-cellular dependencies. The included numerical experiments identify terminal mobility as a key property that must not be disregarded in the radio network planning process, while the deployment of Radio Resource Reservation is shown to be very effective in reducing both call dropping and investment costs.

Chapter 9 concentrates on WLAN performance. In this chapter an integrated packet/flow level modelling approach is presented for analysing flow throughputs and transfer times in IEEE 802.11 WLANs. The packet level model considers persistent data flows and captures the statistical characteristics of the individual packet transmissions according to the CSMA multiple access scheme (distributed packet scheduling), while the flow level model takes into account the system dynamics due to the initiation and completion of data flow transfers. The flow-level model is a PS-based model, which reflects the IEEE 802.11 design principle of distributing the transmission capacity fairly among the active data flows. The integrated model is analytically tractable and yields simple expressions for the expected throughput and flow transfer time. Extensive simulations show that the approximation is very accurate for all considered scenarios. Moreover, the simulation study confirms the attractive property following from our approximation that the expected flow transfer time is insensitive to the flow size distribution (apart from its mean). In a number of numerical experiments, the validated analytical evaluation approach is subsequently used to assess the performance impact of the different link layer access modes and various system and traffic parameters.

An epilogue ends the main body of this monograph with a summary of the key results that have been achieved, and presents an outlook for relevant future research. The complementary CD-ROM contains the entire text of this monograph in PDF format, a full-color presentation of all included figures, all published papers and (conference) presentations, some graphical demonstrations of the simulation software underlying some of the numerical results, as well as some additional more or less informative slide show presentations.

CHAPTER 2

FAIR CHANNEL SHARING IN AN INTEGRATED GSM/HSCSD NETWORK

THE incorporation of High-Speed Circuit-Switched Data in existing speech-oriented GSM networks provided cellular operators with the earliest significant opportunity to offer data services at respectable bit rates. Since cellular operators traditionally have little experience with service integration, the dimensioning and operations of an integrated GSM/HSCSD network impose considerable challenges. In this chapter we develop and analyse a tractable model for the evaluation, comparison and tuning of a generic class of channel sharing policies, that offers great potential to assist the GSM/HSCSD network operators in providing high-quality data services in a resource efficient manner.

The outline of this chapter is as follows. Section 2.1 discusses the relevant literature, followed by a statement of this chapter's contribution in Section 2.2. An overview of HSCSD is provided in Section 2.3. The mathematical framework is set in Section 2.4, which includes the assumptions regarding call characteristics and call handling procedures, and thus defines the studied general class of channel sharing policies. An extensive Markovian performance analysis of the policies in this class is provided in Sections 2.5 and 2.6, where besides some basic performance measures that can be directly derived from the applied Markov chain's equilibrium distribution (Section 2.5), a conditional expected sojourn time analysis is presented for data calls (Section 2.6). Subsequently, four different channel sharing policies that fit within the general class are described in Section 2.7, while a numerical evaluation of these policies is given in Section 2.8. Section 2.9 ends this chapter with some concluding remarks.

2.1. LITERATURE

Aside from a review of the literature that concentrates specifically on the performance of GSM/HSCSD networks, in this section we also present an extensive overview of the relevant literature on the stochastic analysis of Processor Sharing and integrated services models, as this forms the basis underlying the performance analyses presented in this and subsequent chapters.

2.1.1. SERVICE INTEGRATION IN GSM/HSCSD NETWORKS

Few papers have been published that present a performance analysis of an integrated GSM/HSCSD network. In [42, 43] three different channel allocation policies are presented, which are evaluated either by simulation or brute force Markov chain analysis. With their *maximum capacity* policy, an HSCSD call requests a fixed number of channels and is blocked if the requested number of channels is not available. Under the *no rate adaptation* policy an admitted HSCSD call grabs as many channels as there are available, up to its technical maximum transfer capability, and the channel assignment remains fixed for the duration of a call. Under the *rate adaptation* policy a call can grab additional channels as they become available, up to its technical maximum. In the policies proposed in [42, 43] data calls are never downgraded in their assignments, e.g. in order to support a newly originating speech (or data) call. As we will demonstrate, not only upgrading of channel assignments is desirable in order to optimise channel utilisation and delay performance, but also downgrading is essential, primarily to keep the speech call blocking probability low, which is expected to be of great concern to a mobile network operator freshly entering the data market.

[119] presents an evaluation of three types of channel allocation policies in a GSM/HSCSD network. Besides the *maximum capacity* and *no rate adaptation* policies, the authors also propose a ‘soft’ policy where the *fixed* number of channels that are assigned to a new HSCSD data call, and hence also the Call Admission Control rule, is dynamically adjusted based on the blocking statistics. The policies are evaluated by a network simulation with 64 cells, including user mobility. The primary drawback of this study is that the traffic generated consists solely of HSCSD data calls and thus excludes GSM speech calls, while in a realistic network the scarce cell capacity is to be shared between both service types.

Finally, in [14] a single cell in a GSM/HSCSD network is studied serving speech, video and data calls, evaluating a proposed capacity sharing policy in terms of resource utilisation and blocking probabilities.

2.1.2. PROCESSOR SHARING MODELS

An extensive body of literature exists for $M/G/1/PS$ systems with *fixed* service capacity, initiated in [130, 131] for exponentially distributed service requirements ($M/M/1/PS$ queue). There the expected sojourn time in the 'egalitarian' Processor Sharing model conditional on the service requirement was shown to be proportional to the service requirement, while the conditional sojourn time distribution was derived in [52]. The above-mentioned proportionality result was extended to the $M/G/c/PS$ queue in [203, 204]. For $c = 1$ it is a well-known fairness result that the conditional expected sojourn time of a call with a given service requirement x, is equal to $x/(1-\rho)$, where ρ denotes the offered traffic load. As the conditional expected sojourn time depends on the service requirement distribution only through its first moment, also the expected sojourn time is insensitive to the shape of the service requirement distribution (see also [132]). The sojourn time distribution for the $M/G/1/PS$ queue has been derived in [20, 178, 226]. Recently, reliable *prediction* methods for the conditional sojourn time expectation and variance have been developed in [220], given various degrees of available system state information.

Among the generalisations of the 'classical' $M/G/1/PS$ queue, Cohen [54] extends the above results to a general class of networks with a Generalised Processor Sharing service discipline, where the service rate is an arbitrary function of the number of calls in the system. In particular, the (conditional) expected sojourn times in the GPS model also possess the proportionality and insensitivity properties mentioned above. As another generalisation of the standard PS model, the integration of multiple service priority *classes* is considered in the Discriminatory (or weighted) Processor Sharing model, for which class-specific expected sojourn times conditional on the service requirement are derived in [81, 131]. The versatility of the DPS service discipline is paid for by the loss of some of the 'nice' properties of the egalitarian PS systems, e.g. the dependence of the expected sojourn times in the DPS model on the entire service requirement distribution rather than on its first moment alone. Other references that concentrate on QOS differentiation using a DPS-type service discipline include [21, 31, 168]. In [21, 168] exact closed-form expressions and useful approximations are presented for the expected sojourn times of prioritised and best effort calls, respectively. In [31] the relative impact of DPS-type and strict priority queueing schemes on the experienced QOS is assessed for access and backbone links integrating high- and low-priority data traffic and for different degrees of call size variability.

As a final reference on single service PS models, the analysis in the present chapter makes use of the approaches and results reported in [9], which presents a sojourn time analysis for data calls served in a Processor Sharing system with a limited number of service positions, that is fed by a First-In First-Out queue.

Refer to [19, 227, 228] for a more extensive overview of the literature on Processor Sharing systems.

2.1.3. INTEGRATED SERVICES MODELS

In contrast with the references on PS models reviewed above, investigations of integrated services models consider different *types* of services sharing a common resource, and typically concentrate on an integration of loss models for *stream* traffic (e.g. speech telephony) and PS models for *elastic* traffic (e.g. data transfers). As the stream traffic is often assigned some degree of priority in the resource sharing, and elastic traffic flexibly and fairly shares a consequently varying amount of remaining capacity, the elastic QOS analysis in integrated services models typically involves a PS model in a random environment, which severely complicates the analysis (see e.g. [172, Chapter 6]). Moreover, the nice properties of the fixed capacity PS models no longer hold. Only very recently, performance studies of systems with varying service capacity have been published. Although such studies are generally carried out in the wireline context of TCP/IP or ATM networks, the developed models and analyses may be equally applicable in a wireless setting such as second-generation GSM/GPRS networks.

In [175] an analytical comparison of segregated and integrated policies is presented for link sharing in an integrated services TCP/IP (Internet Protocol) or ATM (Asynchronous Transfer Mode) network, where elastic data calls share the varying capacity that is left idle by the prioritised stream calls in a PS fashion. Aside from some basic performance measures that are derived from the equilibrium distribution of the considered stochastic processes, a conditional sojourn time analysis is presented for elastic calls of a given size. A more extensive treatment of this work can be found in

[173, 174]. An analytical performance evaluation of CBR/VBR (Constant/Variable Bit Rate: stream) and ABR/UBR (Available/Unspecified Bit Rate: elastic) traffic sharing an ATM transmission link is presented in [6], where elastic traffic that cannot be served immediately upon admission may be temporarily stored in an access queue for later service. Both an unconditional and conditional elastic call sojourn time analysis is presented, but the latter only for the case with no access queue. A similar model is considered in [5], where stream traffic comes in multiple priority classes to differentiate in call blocking. The (unconditional) stream and elastic call performance is derived using matrix-geometric techniques. For a generic system integrating stream and elastic services, [61] recognises that the stream and elastic traffic dynamics typically occur at distinct time scales and derives distribution-insensitive performance approximations from a quasi-stationary analysis. In a similar model, [17] presents a range of useful approximations and bounds for the elastic QOS, differentiating between a regime of uniform stability, i.e. where regardless of the number of stream calls present, the remaining elastic service capacity suffices to handle the elastic traffic load, and the alternate regime of local instability. In [190] approximations for throughput and blocking performance are derived for processor sharing models integrating two types of elastic services.

2.2. CONTRIBUTION

An extensive performance analysis is presented of a class of fair channel sharing policies in an integrated GSM/HSCSD network serving speech and data calls. Under any policy in the studied class, active data calls are guaranteed a minimum channel assignment, and may be granted the use of additional channels, until these are claimed by new speech or data calls. Data calls can be up- or downgraded in their channel assignment, but only at times of a speech or data call arrival or termination. Admitted data calls that cannot start transfer immediately are queued. Among the principal contributions of this chapter we present a *unified framework* for channel sharing policies, an explicit *analysis* of the (conditional) expected sojourn time (access time *plus* transfer time) of data calls, as well as an extensive *numerical evaluation* of four proposed channel sharing policies. In the presented performance analysis, the data service dynamics are modelled as a queue in a random environment (see e.g. [172, 174]), in the sense that from the perspective of the data service, the environment determining both call admission and channel assignment is defined by the random evolution of the presence of speech calls. In contrast with the general perception of a queue in a random environment, in our case a *mutual* influence between the speech and data service exists.

2.3. HIGH-SPEED CIRCUIT-SWITCHED DATA

GSM and HSCSD share the same radio resources, consisting of a number of *physical channels* that are generated according to a hybrid FD/TDMA scheme, as explained in Chapter 1. A subset of the physical channels is dedicated to carry control signalling on *logical control channels*, e.g. the broadcast, paging and access grant channels (downlink) or the random access channel (uplink). The remaining physical channels carry actual user information on *logical traffic channels* and can be either dedicated to or dynamically shared by GSM speech and HSCSD data calls.

HSCSD enables the assignment of a bundle of traffic channels to a single data call, thereby enhancing the potential information bit rate that can be offered. According to the specifications [69, 71], a single HSCSD data call can be assigned up to eight full rate traffic channels, i.e. an entire GSM carrier, while the assignment may be up- or downgraded during a call in order to optimise service quality and channel utilisation or support newly arriving GSM or HSCSD calls, respectively. See also Figure 2.1 where four traffic channels are simultaneously used by an HSCSD call.

Bundling of traffic channels requires a new functionality (software) in both the Mobile Station and the Base Station Controller. The *terminal adaptation function* in the MS is in charge of splitting and combining the n data substreams that are carried over n traffic channels, and thus forms the interface between the Terminal Equipment (TE) and the radio interface. Across the radio interface, the *inter-working function* in the BSC performs these operations as an interface between the radio interface and the (Gateway) Mobile Switching Centre ((G)MSC), which in turn forms the gateway to external networks.

In principle an HSCSD terminal can be assigned up to eight traffic channels (an entire frequency carrier) in both up- and downlink. However, the first generation of terminals will consist of a single integrated transceiver, allowing downlink assignments of no more than $\beta_{\text{HSCSD}}^{\max} \leq 4$ traffic channels, since the remaining time slots in a GSM TDMA frame are required for uplink transmissions, signal strength measurements, tuning from up- to downlink frequencies, and vice versa (see Figure 2.2). The (downlink) multislot capability is terminal-dependent and is typically 2, 3 or 4. In case of a mobile-originated data transfer, i.e. where the bulk of the data is transferred

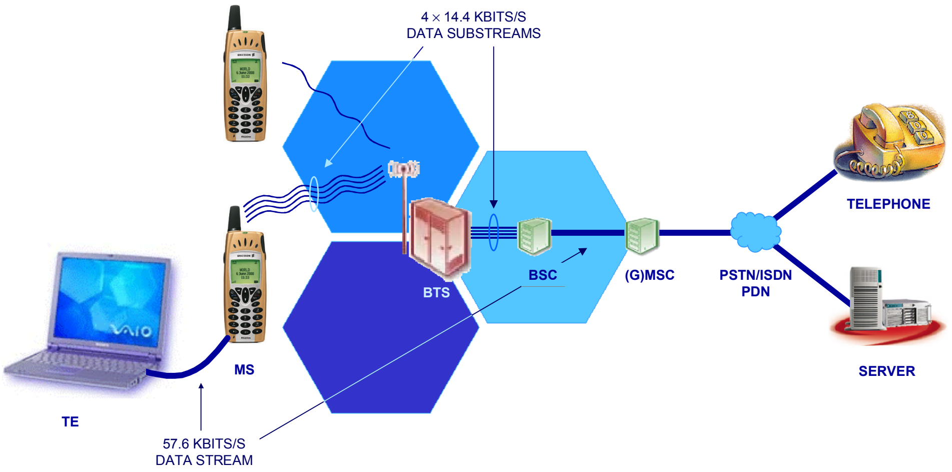


Figure 2.1 GSM/HSCSD network architecture: the illustration shows an example data call at an information bit rate of 4 × 14.4 kbits/s, maintained between an HSCSD terminal and a remote server.

in the uplink, typically no more than two channels can be bundled in the uplink to keep the terminal from becoming too hot.

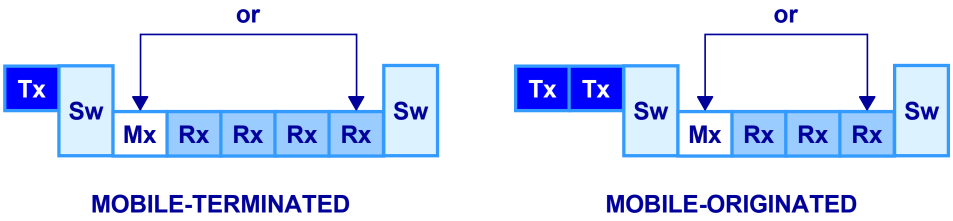


Figure 2.2 Multislot configurations of HSCSD terminals, where 'Tx' refers to transmission, 'Rx' to reception, 'Mx' to signal measurements and 'Sw' to switching between up- and downlink frequencies. The configuration on the left corresponds to a mobile-terminated (dominantly downlink) data transfer, while the configuration on the right corresponds to a mobile-originated (dominantly uplink) data transfer.

A GSM network operator offering HSCSD services to its customers must implement new radio resource management algorithms to optimise resource efficiency and service quality. In this chapter we concentrate on the performance of different Call Admission control and channel assignment policies in a GSM/HSCSD network, in terms of speech and data call blocking probability, data call delays and channel utilisation.

2.4. MODEL

In this section we define the framework for our performance analysis. This framework consists of two principal parts, the assumed service characteristics that specify the traffic model, and the system model, i.e. the cell capacity and the call handling procedures.

2.4.1. TRAFFIC MODEL

Consider a single cell in a GSM/HSCSD network, serving circuit-switched speech and data calls, whose arrival processes and (nominal) call holding time distributions are assumed to be mutually independent.

SPEECH SERVICE: Speech calls arrive according to a Poisson process with arrival intensity λ_{speech} calls per second. An admitted speech call is served with a single dedicated traffic channel for its entire duration, assumed to be exponentially distributed with mean $1/\mu_{\text{speech}}$. The speech traffic load is denoted $\rho_{\text{speech}} \equiv \lambda_{\text{speech}}/\mu_{\text{speech}}$.

DATA SERVICE: Data calls arrive according to a Poisson process with arrival intensity λ_{data} calls per second. A data call is assumed to be the downlink transfer of a file with an exponentially distributed size. We assume an information bit rate of r_{data} kbits/s per traffic channel, and express the data call size in units of r_{data} kbits, so that both the data call size and the data call holding time given the exclusive use of a single traffic channel, are exponentially distributed with mean denoted $1/\mu_{\text{data}}$ (in r_{data} kbits or seconds, respectively). The data traffic load is given by $\rho_{\text{data}} \equiv \lambda_{\text{data}}/\mu_{\text{data}}$. The number of traffic channels that can be assigned to a data call must be between the requested minimum $\beta^{\min}$ and the technical maximum $\beta^{\max}_{\text{HSCSD}}$, with $\beta^{\min} \leq \beta^{\max}_{\text{HSCSD}}$. Since in TDMA-based GSM, a single carrier frequency is time-sliced into eight physical channels, $\beta^{\min}, \beta^{\max}_{\text{HSCSD}} \in \{1, 2, \cdots, 8\}$ must hold. When multiple traffic channels are assigned to a data call, we neglect the technical requirement that these channels must be on the same carrier frequency, for reasons of analytical tractability. As in most cases reshuffling of calls on the available traffic channels can satisfy this technical constraint, the assumption is not as restrictive as it may seem.

2.4.2. SYSTEM MODEL

Denote the cell capacity, i.e. the number of traffic channels available in the considered cell, with C_{total}. A *channel sharing policy* prescribes how the channel pool is shared by speech and data calls, and specifies both a Call Admission Control and a channel assignment policy.

In this chapter we will study the performance of a specific class of channel sharing policies. A common characteristic of all policies in this class is that at a given time all active data calls share the available channel capacity *fairly*, in our model defined such that at any given time the available resources are distributed evenly over the present data calls. Under each such channel sharing policy, the evolution of the system can be modelled as a two-dimensional continuous-time Markov chain $(S(t), D(t))_{t\geq 0}$, where $S(t)$ and $D(t)$ are defined as the number of speech and data calls, respectively, that is present at time t. The system states are denoted (s, d). The data calls present in the system are either *active*, i.e. in transfer on one or more traffic channels, or *queued*, in case no channels can be assigned immediately upon admission. A queued call becomes active once sufficient resources are or can be freed (by downgrading above-guaranteed assignments to other active data calls) to provide the minimum requested channel assignment β^{min}, and it remains active until its transfer is completed and the call terminates. Hence an active data call can never be pushed back into the access queue.

A channel sharing policy in the given class is characterised by three basic functions. Firstly, $\beta(s, d)$ is the expected number of channels that are assigned to each active data call in state (s, d), i.e. the total number of traffic channels assigned to data calls divided by the number of *active* data calls. By convention, $\beta\,(s, d) = 0$ if there are no active data calls. The circuit-switched character of the HSCSD service allows only an integer number of channels to be assigned to an active data call. In the considered model, the proposed *fair* channel sharing policies (re)assign, at each call arrival or termination event, the traffic channels available for data transfer as follows. Each active data call receives a basic assignment of $\underline{\beta}(s, d) \equiv \lfloor \beta(s, d) \rfloor$ channels which is the maximum assignment that can be uniformly awarded, while the remaining channels are randomly distributed over the active data calls, so that some fortunate data calls obtain an additional channel and are assigned $\overline{\beta}(s,d) \equiv \lceil \beta(s,d) \rceil = \underline{\beta}(s,d)+1$ channels. Hence an active data call is assigned either $\underline{\beta}(s,d)$ or $\overline{\beta}(s,d)$ channels with respective probabilities $\underline{p}(s,d) \equiv \lceil \beta(s,d) \rceil - \beta(s,d)$ and $\overline{p}(s,d) \equiv 1 - \underline{p}(s,d) = \beta(s,d) - \lfloor \beta(s,d) \rfloor$. It is then intuitively clear and easily verified that the expected number of channels assigned to a data call equals $\beta(s,d)$, i.e.

$$\underline{p}(s,d)\,\underline{\beta}(s,d) + \overline{p}(s,d)\,\overline{\beta}(s,d) = \beta(s,d). \tag{2.1}$$

Secondly, the maximum number of data calls that can be allowed in the cell when there are s speech calls present, is denoted $d_{\max}(s)$, whose specific form depends on the selected channel sharing policy. Incorporated in $d_{\max}(s)$ is the possibility of queueing data calls (see also Figure 2.3).

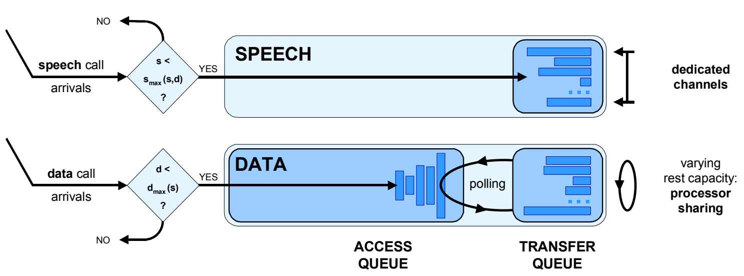


Figure 2.3 The considered system model integrates speech and data calls according to specific call handling procedures. The *access queue* stores data call requests that cannot be served immediately upon admission. Once sufficient capacity is or can be freed, a data call in the access queue is polled into the *transfer queue* and the requested data volume is transferred.

An admitted data call that cannot start service immediately is held in a fixed-size First-In First-Out access queue that can store up to Q_a requests, until the system has evolved to a state where sufficient capacity is freed to serve the call and the call is polled into the transfer queue. We stress that freed capacity should indeed be immediately assigned to a queued data call, if possible, rather than allowing the active calls to be served faster, as this reflects the rational objective to handle call requests as quickly as possible, while still respecting the minimum QOS requirements. An alternative policy could be to implement a dynamic access queue, in the sense that the admission of a data call request into the access queue is based on the expected access time, conditional on the actual system state upon arrival. Such a policy can still be evaluated with the analytical techniques presented below. It is noted that although the GSM/HSCSD standards do allow queueing of circuit-switched calls (see [77]), not all manufacturers support this feature. In practice, a fresh (or handover) call request that cannot be assigned a traffic channel immediately can be put in a BSC queue in the hope that sufficient capacity is freed before a time-out occurs. Effectively, this means that the admit/reject decision is postponed for a few seconds. In our model the queueing feature is implemented for HSCSD data calls only, as a higher delay tolerance is expected from the corresponding users. Still, the value of Q_a must be small in order to appropriately mimic the typically small time-out value that characterises the more practical policy, which is unfortunately analytically rather intractable in its exact form. As will be demonstrated in Section 2.8, the optimal size of Q_a depends on multiple factors, e.g. the cell capacity C_{total}, the data traffic load ρ_{data}, and even on the average call size $1/\mu_{\text{data}}$, since for a given data traffic load, a smaller average call size implies that more data calls can be queued and still be served within the allowed call setup time.

In order to define an equivalent function for the maximum number of admissible speech calls, we note that the admission of a newly arriving speech call depends on the actual speech/data call configuration, i.e. on the system state (s, d), rather than on d alone. The rationale behind this is that the number of active and queued data calls is determined by both s and d, and hence the number of available traffic channels as well. Note hereby that it is possible for a data call to be queued even when some traffic channels are idle, either because these idle channels are reserved for speech calls, or because the number of channels that are (or can be made) available is less than $\beta^{\min}$. As a consequence, (s, d) determines whether or not a new speech call can be admitted, and thus the third function to be specified is denoted $s_{\max}(s, d)$, defined such that $s_{\max}(s, d) - s$ is the maximum number of speech calls that can still be admitted, starting from state (s, d).

Consider a simple illustrative example with $C_{\text{total}} = 2$, $\beta^{\min} = \beta^{\max}_{\text{HSCSD}} = 2$, $Q_a = 1$, and full sharing of traffic channels between speech and data calls. The corresponding state space is depicted in Figure 2.4, where for each admissible state (s, d), $s_{\max}(s, d)$ and $d_{\max}(s)$ are given. The block arrows indicate the possible state transitions. Note that $s_{\max}(0, 1) = 0$ since there is one active data call, occupying both channels, so that a newly arriving speech call must be blocked. Compare this with state $(1, 1)$ with one active speech call and one queued data call, where $s_{\max}(1,1) = 2$, since there is one idle channel that may be assigned to a new speech call. Apparently, we need to know the number of speech calls s in the system, to know how many of the d data calls are active or queued, which in turn determines how many more speech calls can be admitted.

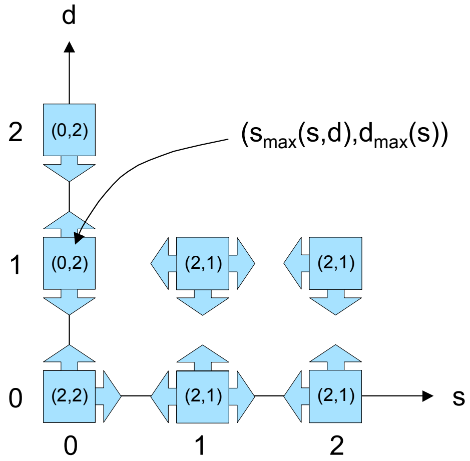


Figure 2.4 Illustration of state space, $s_{\max}(s,d)$ and $d_{\max}(s)$: the example indicates why the speech calls' CAC threshold $s_{\max}$ depends on both s and d.

Two additional parameters that will be useful in the analysis can be derived from the basic functions. Denote with $s_{\max} \equiv s_{\max}(0,0)$ and $d_{\max} \equiv d_{\max}(0)$ the absolute maximum number of speech and data calls the system can support, respectively.

These three basic functions are sufficient to specify a Call Admission Control policy. If at time t a newly originating speech call sees the system in state (s,d), it is admitted if and only if $s < s_{\max}(s,d)$, i.e. if and only if there is still a spare channel to serve the speech call (there is no access queue for speech calls). Similarly, a new data call is admitted if and only if $d < d_{\max}(s)$. If $Q_a > 0$ then $d = d_{\max}(s)$ if and only if the access queue is full. All blocked calls are cleared from the system. Furthermore, these functions fully specify a *channel assignment* policy as well. In state (s,d), each of the s speech calls is active and assigned a single channel. Of the d data calls present, $d_t(s,d) \equiv \min\{d, d_{\max}(s) - Q_a\}$ are in active transfer, fairly sharing a total of $d_t(s,d)\,\beta(s,d)$ channels, while the remaining $d_a(s,d) \equiv d - d_t(s,d)$ data calls are 'stored' in the access queue.

2.4.3. PERFORMANCE MEASURES

The system level performance is assessed in terms of the *expected channel utilisation.* The GOS of the speech and data service is expressed by the *call blocking probabilities*, while the experienced data QOS is primarily given by the *(conditional) expected sojourn time.*

2.5. BASIC PERFORMANCE ANALYSIS

In this section we capture the traffic and system model ingredients specified above in a Markov chain model. Subsequently, a number of performance measures is specified that can be directly derived from the Markov chain's equilibrium distribution.

2.5.1. MARKOV CHAIN

Under any channel sharing policy in the considered class, the evolution of the system can be described by an irreducible two-dimensional continuous-time Markov chain $(S(t), D(t))_{t\geq 0}$, with states denoted (s, d). The state space of the Markov chain is given by

$$\mathbb{S} \equiv \{(s,d) \in \mathbb{N}_0 \times \mathbb{N}_0 : s \leq s_{\max}(s,d) \text{ and } d \leq d_{\max}(s)\}\,.$$

Ordering the state space lexicographically in (s, d), the infinitesimal generator is given by

$$\mathcal{Q} \equiv \begin{pmatrix} \mathcal{C}_0 & \mathcal{A}_0 & \mathcal{O} & \cdots & \cdots & \mathcal{O} \\ \mathcal{B}_1 & \mathcal{C}_1 & \mathcal{A}_1 & \mathcal{O} & \ddots & \vdots \\ \mathcal{O} & \mathcal{B}_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathcal{A}_{s_{\max}-2} & \mathcal{O} \\ \vdots & \ddots & \mathcal{O} & \mathcal{B}_{s_{\max}-1} & \mathcal{C}_{s_{\max}-1} & \mathcal{A}_{s_{\max}-1} \\ \mathcal{O} & \cdots & \cdots & \mathcal{O} & \mathcal{B}_{s_{\max}} & \mathcal{C}_{s_{\max}} \end{pmatrix} \in \mathbb{R}^{|\mathbb{S}|} \times \mathbb{R}^{|\mathbb{S}|},$$

with $|\mathbb{S}| = \sum_{s=0}^{s_{\max}} (d_{\max}(s) + 1)$.

The *super-diagonal* blocks $\mathcal{A}_s$ generate speech call arrival events, for $s = 0, \cdots, s_{\max} - 1$. $\mathcal{A}_s$ is of dimension $(d_{\max}(s)+1) \cdot (d_{\max}(s+1)+1)$ and has entries $\mathcal{A}_s(d,d) \equiv \lambda_{\text{speech}} 1\{s < s_{\max}(s,d)\}$ for $d = 0, \cdots, d_{\max}(s+1)$, and $\mathcal{A}_s(d,d') \equiv 0$ for $d \neq d'$, where the indicator function $1\{\cdot\}$ returns 1 if the argument event is true, and 0 otherwise. Note that the inclusion of the factor $1\{s < s_{\max}(s,d)\}$ is necessary to cope with a situation as described in Section 2.4.2, where a state $(s+1,d)$ is in principle admissible, yet it cannot be reached from state (s,d). The *sub-diagonal* blocks $\mathcal{B}_s$ generate speech call termination events, for $s = 1, \cdots, s_{\max}$. $\mathcal{B}_s$ is of dimension $(d_{\max}(s)+1) \times (d_{\max}(s-1)+1)$ and has entries $\mathcal{B}_s(d,d) \equiv s\,\mu_{\text{speech}}$ for $d = 0, \cdots, d_{\max}(s)$, and $\mathcal{B}_s(d,d') \equiv 0$ for $d \neq d'$. Finally, the square blocks $\mathcal{C}_s$ on the *diagonal* generate data call arrival and termination events, for $s = 0, \cdots, s_{\max}$. $\mathcal{C}_s$ is of dimension $(d_{\max}(s)+1) \times (d_{\max}(s)+1)$ and has entries $\mathcal{C}_s(d-1,d) \equiv \lambda_{\text{data}}$ and $\mathcal{C}_s(d,d-1) \equiv \beta(s,d)\,d_t(s,d)\,\mu_{\text{data}}$ for $d = 1, \cdots, d_{\max}(s)$. Furthermore, the diagonal entries of $\mathcal{C}_s$ are such that the entries of each row of $\mathcal{Q}$ sum up to 0. All other entries of $\mathcal{C}_s$ are equal to zero.

Since the finite state space Markov chain is irreducible, there is a unique probability vector $\boldsymbol{\pi}$ that satisfies the system of balance equations (e.g. [213])

$$\boldsymbol{\pi}\mathcal{Q} = \mathbf{0},$$

with $\mathbf{0}$ the vector with all entries zero, and $\boldsymbol{\pi}$ lexicographically ordered in the system states $(s,d) \in \mathbb{S}$.

Remark 2.1 While the applied ordering of the state variables s and d is chosen in accordance with [173, 174, 175], a reverse ordering of the state variables is also possible, requiring a corresponding modification of the infinitesimal generator. Such an alternative ordering is most convenient for the analysis of a similar system but without call admission control for data traffic, i.e. with $d_{\max} = \infty$, in which case matrix-geometric techniques can be efficiently applied to determine the Markov chain's equilibrium distribution (see e.g. [172]).

2.5.2. BASIC PERFORMANCE MEASURES

A number of performance measures can be directly determined from the equilibrium distribution of the Markov chain. From a system perspective, the resource efficiency yielded by each channel sharing policy can be measured by the *expected channel* *utilisation,*

$$\mathbf{U} \equiv C_{\text{total}}^{-1} \sum_{(s,d)\in\mathbb{S}} (s + \beta(s,d)\, d_t(s,d))\ \pi(s,d).$$

The performance of the channel sharing policies with respect to the speech and data services is primarily measured by the *call blocking probabilities,*

$$\mathbf{P}_{\text{speech}} \equiv \sum_{(s,d)\in\mathbb{S}} 1\{s = s_{\max}(s,d)\}\, \pi(s,d),$$

and

$$\mathbf{P}_{\text{data}} \equiv \sum_{0\le s\le s_{\max}} \pi(s, d_{\max}(s)),$$

using the PASTA (Poisson Arrivals See Time Averages) property (e.g. [224]).

With the expected number of data calls in the access and transfer queues given by $\mathbf{N}_{t,\text{data}} \equiv \sum_{(s,d)\in\mathbb{S}} d_t(s,d)\, \pi(s,d)$ and $\mathbf{N}_{a,\text{data}} \equiv \sum_{(s,d)\in\mathbb{S}} d_a(s,d)\, \pi(s,d)$, respectively, the *expected access* and *transfer times* of a data call are given by

$$\mathbf{T}_{a,\text{data}} \equiv \frac{\mathbf{N}_{a,\text{data}}}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})} \text{ and } \mathbf{T}_{t,\text{data}} \equiv \frac{\mathbf{N}_{t,\text{data}}}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})},$$

respectively, using Little's formula [224]. As a consequence, the *expected sojourn time* of a data call is equal to $\mathbf{T}_{\text{data}} \equiv \mathbf{T}_{a,\text{data}} + \mathbf{T}_{t,\text{data}}$.

Considered from a system's perspective, the *expected (time-average) throughput* experienced by an *active* data call is defined by

$$\mathbf{R}_{\text{data}}^{t} \equiv r_{\text{data}} \sum_{(s,d)\in\mathbb{S}_{+}^{t+}} \left(\frac{\pi(s,d)}{\sum\limits_{(s,d)\in\mathbb{S}_{+}^{t+}} \pi(s,d)} \right) \beta(s,d),$$

where $\mathbb{S}_{+}^{t+} \equiv \{(s,d)\in\mathbb{S} : d_t(s,d) > 0\}$ is the set of states with at least one active data call (see also Figure 2.5 below). The expected (time-average) number of traffic channels assigned to an *active* data call must obviously lie somewhere within the range of technical limitations, i.e. $r_{\mathrm{data}}^{-1}\mathbf{R}_{\mathrm{data}}^{t} \in \left[\beta^{\min}, \beta^{\max}_{\mathrm{HSCSD}}\right]$. Furthermore, it is stressed that the effects of the access time are not incorporated in $\mathbf{R}_{\mathrm{data}}^{t}$, so that it is a meaningful throughput measure only if the access times are typically very small in comparison with the transfer times.

2.6. CONDITIONAL PERFORMANCE ANALYSIS

The expected sojourn time $\mathbf{T}_{\mathrm{data}}$ of a data call can be easily computed, once the steady state probabilities have been determined, as demonstrated above. For data calls we are also interested in the *conditional* expected sojourn time given the system state upon call arrival and/or the call size. The merit of the conditional performance measures is twofold: on the one hand, the demonstrated asymptotic linearity of the conditional expected transfer time in the data call size indicates a *fairness* property, while on the other hand, the conditional expected sojourn times can be fed back to the caller as an *informative* indication of the expected call handling time (provided that information regarding the data call size is indeed available).

It is convenient to partition the state space $\mathbb{S}$ and introduce some additional notation. Denote with $\mathbb{S}_0 \equiv \{(s,d) \in \mathbb{S} : d = 0\}$ the set of states with no data calls, and with $\mathbb{S}_+ \equiv \mathbb{S}\backslash\mathbb{S}_0$ its complement. Further, let $\mathbb{S}_+^{t_0} \equiv \{(s,d) \in \mathbb{S}_+ : d_t(s,d) = 0\}$ denote the subset of $\mathbb{S}_+$ with no active data calls, while $\mathbb{S}_+^{t_+} \equiv \mathbb{S}_+\backslash\mathbb{S}_+^{t_0}$ is its complement within $\mathbb{S}_+$. Lastly, let $\mathbb{S}_+^{a_0} \equiv \{(s,d) \in \mathbb{S}_+ : d_a(s,d) = 0\}$ be the subset of $\mathbb{S}_+$ with no data calls in the access queue, and let $\mathbb{S}_+^{a_+} \equiv \mathbb{S}_+\backslash\mathbb{S}_+^{a_0}$ be its complement in $\mathbb{S}_+$. This partitioning is illustrated in Figure 2.5.

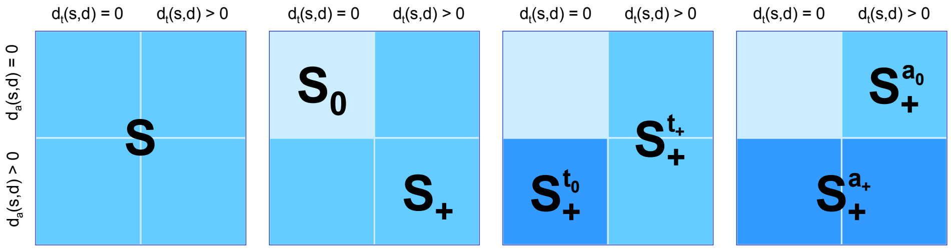


Figure 2.5 Illustration of the state space partitioning required for the conditional performance analysis, based on whether the access and transfer queues are (non-)empty.

For each $(s,d) \in \mathbb{S}_+$ define $\sigma_{s,d}(x)$ as the random sojourn time of an admitted data call of size x, arriving at *given* system state (s,d) with s speech and d data calls (d *includes* the new call), and let $\widehat{\sigma}_{s,d}(x) \equiv \mathbf{E}\{\sigma_{s,d}(x)\}$ denote its expectation with respect to the random process of the system evolution. Then the expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of an admitted data call of size x is given by

$$\mathbf{T}_{\text{data}}(x) = \sum_{(s,d)\in\mathbb{S}_+} \left(\frac{\pi(s,d-1)}{\sum\limits_{(s,d)\in\mathbb{S}_+} \pi(s,d-1)} \right) \widehat{\sigma}_{s,d}(x), \tag{2.2}$$

where

$$\sum_{(s,d)\in\mathbb{S}_+} \pi(s,d-1) = \sum_{\substack{(s,d)\in\mathbb{S} \\ 0\le d<d_{\max}(s)}} \pi(s,d) = 1 - \mathbf{P}_{\text{data}}.$$

Note that $\pi(s,d)/\left(1-\mathbf{P}_{\text{data}}\right)$ is the equilibrium probability that an *accepted* data call finds s speech calls and d *other* data calls in the system, which is stressed to be generally unequal to $\pi(s,d)$ as the PASTA property does not apply to *accepted* calls. Hence

$$\mathbf{T}_{\text{data}} = \int\limits_{x=0}^{\infty} \mathbf{T}_{\text{data}}(x)\, \mu_{\text{data}}\, e^{-x\,\mu_{\text{data}}}\, dx.$$

In order to determine the functions $\widehat{\sigma}_{s,d}(x)$, $(s,d) \in \mathbb{S}_+$, $x \in \mathbb{R}^+$, we may split the expected sojourn time into an access time and a transfer time component, along the lines followed in [9], where a sojourn time analysis is presented for a single server system serving a single type of calls with a PS service discipline, and a limited number of service positions.

In analogy with the sojourn time variables, denote with $\alpha_{s,d}$ the random access time of a newly admitted data call of size x, entering the system in state (s,d), which is obviously independent of the call size, due to the First-In First-Out queueing discipline, and let $\widehat{\alpha}_{s,d} \equiv \mathbf{E}\{\alpha_{s,d}\}$ denote its expectation. Similarly, denote with $\tau_{s_o,d_o}(x)$ the random transfer time of an admitted data call of size x, that starts its transfer in system state $(s_o,d_o) \in \mathbb{S}_+^{t+}$, and let $\widehat{\tau}_{s_o,d_o}(x) \equiv \mathbf{E}\{\tau_{s_o,d_o}(x)\}$ denote its expectation. The final ingredient required to formulate Proposition 2.1 below is $\Psi\left((s,d)\,;(s_o,d_o)\right)$, defined as the probability that a data call entering the system in state $(s, d) \in \mathbb{S}_+$ starts its transfer in state $(s_o, d_o) \in \mathbb{S}_+^{t+}$.

Proposition 2.1 *The conditional expected sojourn time* $\widehat{\sigma}_{s,d}(x)$, $(s, d) \in \mathbb{S}_+$, $x \in \mathbb{R}^+$, *can be expressed as the sum of the expected access time and the expected transfer time, as follows:*

$$\widehat{\sigma}_{s,d}(x) = \widehat{\alpha}_{s,d} + \sum_{(s_o,d_o)\in\mathbb{S}_+^{t+}} \widehat{\tau}_{s_o,d_o}(x)\,\Psi\left((s,d)\,;(s_o,d_o)\right). \tag{2.3}$$

Proof The result immediately follows from conditioning on the system state (s_o, d_o) where the tagged data call starts its transfer. □

The $\widehat{\alpha}_{s,d}$, $(s, d) \in \mathbb{S}_+$, are not required to obtain $\mathbf{T}_{\text{data}}(x)$, which becomes clear when substituting (2.3) in (2.2). Still, the values of $\widehat{\alpha}_{s,d}$ are of interest to get some insight into the variability of the access times in view of the practical restriction on how long a data call request can be queued. Furthermore, a mobile station may be informed of the expected delay until it can start its data transfer.

Now that we have justified the above decomposition of the expected sojourn time, we will derive expressions to calculate the expected access and transfer times, as well as the required transition probabilities. Once these expressions are derived, all ingredients are available to compute the expected conditional sojourn time of a data call, using (2.2) and (2.3).

2.6.1. ACCESS TIME ANALYSIS

Consider the continuous-time Markov chain that results when the data call arrival process in the original Markov chain is turned off upon arrival of a tagged data call. With this modification, the system state explicitly indicates the tagged call's position in the access queue, which enables us to determine its expected access time, given by the time until the access queue is emptied. The modified Markov chain has infinitesimal generator $\widetilde{\mathcal{Q}}$, which is similar to $\mathcal{Q}$, except that all data call arrival rates have been set to zero, i.e. $\lambda_{\text{data}} = 0$, while the diagonal elements are modified accordingly. Hence only the submatrices $\widetilde{\mathcal{C}}_s$ differ from $\mathcal{C}_s$, $s = 0, \cdots, s_{\max}$.

The state space of the modified Markov chain remains $\mathbb{S}$. The analysis conveniently utilises the introduced partitioning of $\mathbb{S}_+$ into the (h.l.) absorbing set $\mathbb{S}_+^{a_0}$of states where the access queue is empty and its (h.l.) transient complement, $\mathbb{S}_+^{a+}$.

As was argued above, the access time of a queued data call in the original Markov chain is independent of the future data call arrival process. Hence $\widehat{\alpha}_{s,d}$ is equal to the expected time it takes for the modified chain to evolve from state (s, d) into any state in the absorbing set $\mathbb{S}_+^{a_0}$, i.e. $\widehat{\alpha}_{s,d}$ is the expected *absorption time* of $\mathbb{S}_+^{a_0}$, starting from state (s, d). Trivially, $\widehat{\alpha}_{s,d} = 0$ for all $(s, d) \in \mathbb{S}_+^{a_0}$ since the tagged call can start its transfer immediately upon arrival. Let the vector

$$\widehat{\boldsymbol{\alpha}}_+^n, \; n \in \mathbb{N}_0,$$

contain the cumulative expected access time $\widehat{\alpha}_{s,d}^n$ after n state transitions from initial state $(s, d) \in \mathbb{S}_+^{a+}$, lexicographically ordered in the (s, d). Then this vector of cumulative expected access times evolves according to the following recursive relation:

$$\widehat{\boldsymbol{\alpha}}_+^{n+1} = \widehat{\boldsymbol{\alpha}}_+^\star + \widetilde{\mathcal{P}}_{++}\, \widehat{\boldsymbol{\alpha}}_+^n, \; n \in \mathbb{N}_0, \text{ with initial condition } \widehat{\boldsymbol{\alpha}}_+^0 = \mathbf{0}, \tag{2.4}$$

where the vector $\widehat{\boldsymbol{\alpha}}_+^\star$ contains the expected access times $\left(\widehat{\alpha}_{s,d}^\star, \; (s, d) \in \mathbb{S}_+^{a+}\right)$ until the next state transition, obtained by conditioning on the possible events:

$$\widehat{\alpha}_{s,d}^\star = \left[\lambda_{\text{speech}}\, 1\left\{s < s_{\max}\right\} + s\, \mu_{\text{speech}} + \beta(s, d)\, d_t(s, d)\, \mu_{\text{data}}\right]^{-1}.$$

$\widetilde{\mathcal{P}}_{++}$ denotes the one-step transition probability matrix of the embedded jump chain that follows the transitions within $\mathbb{S}_+^{a+}$ of the modified continuous-time Markov chain with infinitesimal generator $\widetilde{\mathcal{Q}}$, i.e.

$$\widetilde{\mathcal{P}}_{++}((s, d); (s', d')) = \begin{cases} \dfrac{\widetilde{\mathcal{Q}}((s, d); (s', d'))}{-\widetilde{\mathcal{Q}}((s, d); (s, d))} & \text{if } (s, d) \neq (s', d'), \\ \\ 0 & \text{otherwise,} \end{cases}$$

for all $(s, d), (s', d') \in \mathbb{S}_+^{a+}$. Note that probability matrix $\widetilde{\mathcal{P}}_{++}$ is substochastic since it excludes all possible state transitions from $\mathbb{S}_+^{a+}$ to $\mathbb{S}_+^{a_0}$, and has dimensions $|\mathbb{S}_+^{a+}| \times$

$|\mathbb{S}_+^{a+}|$. Finally, the initial condition of the recursive relation simply reflects that the cumulative expected access time of the tagged data call is initialised to zero when it enters the system.

We can now formulate the following proposition.

Proposition 2.2 *The expected access time of a data call which enters the system in state (s, d) in $\mathbb{S}_+^{a+}$ is contained in the vector $\widehat{\boldsymbol{\alpha}}_+$, given by*

$$\widehat{\boldsymbol{\alpha}}_+ \quad = \left(\mathcal{I} - \widetilde{\mathcal{P}}_{++}\right)^{-1} \widehat{\boldsymbol{\alpha}}_+^\star .$$

Proof The expected access time of a data call which enters the system in state (s, d) is given in the limit value of the cumulative expected access time vector of the modified Markov chain, governed by recursive relation (2.4), and is given by

$$\widehat{\boldsymbol{\alpha}}_+ = \lim_{n\to\infty} \widehat{\boldsymbol{\alpha}}_+^n .$$

The limit value follows from solving the linear system of balance equations provided by (2.4):

$$\widehat{\boldsymbol{\alpha}}_+ = \widehat{\boldsymbol{\alpha}}_+^\star + \widetilde{\mathcal{P}}_{++}\, \widehat{\boldsymbol{\alpha}}_+ .$$

The convergence of the cumulative expected access times is due to the transience of $\mathbb{S}_+^{a+}$ and the fact that no further costs are incurred in the absorbing set $\mathbb{S}_+^{a_0}$. Indeed, since $\widetilde{\mathcal{P}}_{++}$ is a substochastic matrix representing transient states, $\widetilde{\mathcal{P}}_{++}^n \to 0$, which implies that all of the eigenvalues of $\widetilde{\mathcal{P}}_{++}$ have absolute values strictly less than 1. Hence the eigenvalues of $\mathcal{I} - \widetilde{\mathcal{P}}_{++}$ are all non-zero, and thus the matrix is indeed non-singular and its inverse $\left(\mathcal{I} - \widetilde{\mathcal{P}}_{++}\right)^{-1}$ exists, which concludes the proof. □

2.6.2. TRANSITION PROBABILITIES

We now compute the probability matrix Ψ containing the probabilities $\Psi((s, d)\,; (s_o, d_o))$ that a call entering the system in state $(s, d) \in \mathbb{S}_+$ starts its transfer in state $(s_o, d_o) \in \mathbb{S}_+^{t+}$. Hence Ψ is of dimension $|\mathbb{S}_+| \times |\mathbb{S}_+^{t+}|$. Some of the entries of Ψ are readily determined. For instance, $\Psi\left((s, d)\,; (s, d)\right) = 1$ for $(s, d) \in \mathbb{S}_+^{a_0}$, since the arriving data call can immediately start its transfer.

In order to compute Ψ, we augment the state space $\mathbb{S}$ of the original Markov chain, i.e. *including* data call arrivals and generated by $\mathcal{Q}$, with an extra dimension. Denote with $N(t)$ the location in the access queue of a tagged data call at time t, and let $N(t) = 0$ if the data call is no longer queued, i.e. it is either in transfer or already fully processed. The augmented Markov chain is denoted $(N(t), S(t), D(t))_{t\geq 0}$, with states (n, s, d). For a given channel sharing policy, the state space of the augmented Markov chain is given by

$$\mathbb{S}^\bullet \equiv \{(n, s, d) \in \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0 : n \leq d_a(s, d) \text{ and } (s, d) \in \mathbb{S}\}.$$

The state space is partitioned into an absorbing subset, $\mathbb{S}^\bullet_0 \equiv \{(n, s, d) \in \mathbb{S}^\bullet : n = 0\}$, in which the tagged call is no longer queued, and its transient complement, $\mathbb{S}^\bullet_+ \equiv \mathbb{S}^\bullet \backslash \mathbb{S}^\bullet_0$. We further modify the augmented Markov chain, by reducing all rates out of any state in $\mathbb{S}^\bullet_0$ to zero, thereby enforcing that each such state becomes absorbing. The modified chain is a continuous-time Markov chain with state space $\mathbb{S}^\bullet$, that consists of $|\mathbb{S}^\bullet_0| = |\mathbb{S}|$ absorbing states and one transient class $\mathbb{S}^\bullet_+$.

Denote with $\mathcal{P}^\bullet$ the one-step transition probability matrix of the embedded jump chain that follows the transitions of the augmented continuous-time Markov chain, with entries given by

$$\mathcal{P}^\bullet((n, s, d); (n', s', d')) =$$

$$= \begin{cases} 1 & \text{if } n = 0 \text{ and } (n', s', d') = (n, s, d), \\[2ex] \dfrac{\mathcal{Q}((s, d); (s', d'))}{-\mathcal{Q}((s, d); (s, d))} & \text{if } n > 0 \text{ and } \begin{cases} (n', s', d') \in \{(n, s{+}1, d), (n, s, d{+}1), \\ \qquad\qquad (n{-}1, s, d{-}1)\}, \text{ or} \\ (n', s', d') = (n, s{-}1, d) \text{ and} \\ \qquad\qquad d_t(s{-}1, d) = d_t(s, d), \text{ or} \\ (n', s', d') = (n{-}1, s{-}1, d) \text{ and} \\ \qquad\qquad d_t(s{-}1, d) = d_t(s, d) + 1 \end{cases} \\[2ex] 0 & \text{otherwise,} \end{cases}$$

for all $(n, s, d), (n', s', d') \in \mathbb{S}^\bullet$. Clearly, all transitions that have non-zero probability of occurrence correspond to speech call arrivals, speech call terminations, data call arrivals, and data call terminations, respectively. Note that n changes at each data call termination event, as well as at the termination of a speech call whose released channels enable a queued data call to become active. $\mathcal{P}^\bullet$ can be written in the form

$$\mathcal{P}^\bullet \equiv \begin{pmatrix} \mathcal{I} & \mathcal{O} \\ \mathcal{P}^\bullet_{+0} & \mathcal{P}^\bullet_{++} \end{pmatrix},$$

where $\mathcal{I}$ is the identity matrix, $\mathcal{O}$ is the null-matrix, and $\mathcal{P}^\bullet_{+0}$ and $\mathcal{P}^\bullet_{++}$ are sub-stochastic submatrices of $\mathcal{P}^\bullet$ corresponding to the transitions from $\mathbb{S}^\bullet_+$ to $\mathbb{S}^\bullet_0$ and $\mathbb{S}^\bullet_+$, respectively.

Proposition 2.3 *The probability $\Psi\left((s, d)\,; (s_o, d_o)\right)$ that a call entering the system in state $(s, d) \in \mathbb{S}_+$ (d includes the new call) starts its transfer in state $(s_o, d_o) \in \mathbb{S}^{t+}_+$, is equal to the element $\Psi^\bullet\left((d_a(s, d), s, d)\,;\;\; (0, s_o, d_o)\right)$ of the probability matrix $\Psi^\bullet$ given by*

$$\Psi^\bullet = \left(\mathcal{I} - \mathcal{P}^\bullet_{++}\right)^{-1} \mathcal{P}^\bullet_{+0}.$$

The matrix $\Psi^\bullet$ is of dimension $|\mathbb{S}^\bullet_+| \times |\mathbb{S}^\bullet_0|$ and contains the probabilities that the augmented chain, starting in any transient state in $\mathbb{S}^\bullet_+$ is eventually absorbed in any of the recurrent states in $\mathbb{S}^\bullet_0$.

Proof It is obvious that the probability $\Psi\left((s, d)\,; (s_o, d_o)\right)$ that a call entering the original system in state $(s, d) \in \mathbb{S}_+$ starts its transfer in state $(s_o, d_o) \in \mathbb{S}^{t+}_+$ is equal to the probability $\Psi^\bullet\left((d_a(s, d), s, d)\,; (0, s_o, d_o)\right)$ that the augmented process, starting in state $(d_a(s, d), s, d) \in \mathbb{S}^\bullet_+$ is eventually absorbed in state $(0, s_o, d_o) \in \mathbb{S}^\bullet_0$. Hence we only need to show that the probability matrix $\Psi^\bullet$ can indeed be calculated as stated in the proposition.

Consider the augmented chain in transient state $(n, s, d) \in \mathbb{S}^\bullet_+$. Conditioning on the first transition out of (n, s, d) yields, for any $(0, s_o, d_o) \in \mathbb{S}^\bullet_0$,

$$\Psi^\bullet\left((n, s, d)\,; (0, s_o, d_o)\right) = \mathcal{P}((n, s, d); (0, s_o, d_o))$$

$$+ \sum_{(n',s',d')\in\mathbb{S}_{+}^{\bullet}} \mathcal{P}((n,s,d);(n',s',d'))\,\Psi^{\bullet}\left((n',s',d');(0,s_o,d_o)\right).$$

In matrix form, this can be formulated as

$$\Psi^{\bullet} = \mathcal{P}_{+0}^{\bullet} + \mathcal{P}_{++}^{\bullet}\Psi^{\bullet}.$$

Note that $\mathcal{P}_{++}^{\bullet}$ is a substochastic matrix representing transient states, so that $\left(\mathcal{P}_{++}^{\bullet}\right)^{n} \to 0$, which implies that all of the eigenvalues of $\mathcal{P}_{++}^{\bullet}$ have absolute values strictly less than 1. Hence the eigenvalues of $\mathcal{I} - \mathcal{P}_{++}^{\bullet}$ are all non-zero, and thus the matrix is non-singular and its inverse $\left(\mathcal{I} - \mathcal{P}_{++}^{\bullet}\right)^{-1}$ exists, which concludes the proof. □

2.6.3. TRANSFER TIME ANALYSIS

We now focus on the conditional expected transfer time $\widehat{\tau}_{s,d}(x)$ of a tagged active data call of length $x \geq 0$, starting its transfer in the presence of s speech and d data calls, $(s,d) \in \mathbb{S}_{+}^{t+}$, including itself and all queued data calls. Data call length x is expressed in units of r_{data} kbits so that it takes x seconds to transfer a file of length x on a single dedicated traffic channel. In the following an explicit expression for vector $\widehat{\boldsymbol{\tau}}(x) = \left(\widehat{\tau}_{s,d}(x),\,(s,d) \in \mathbb{S}_{+}^{t+}\right)$ is derived. The presented transfer time analysis is based on similar results reported in [173, 175]. It is noted that apart from the conditional *expected* transfer times, a closed-form expression for the Laplace-Stieltjes transform of the *distribution* of $\tau_{s,d}(x)$, $(s,d) \in \mathbb{S}_{+}^{t+}$, can be obtained along similar lines (see also Chapter 3 and [173]).

We need to make another modification to the original Markov chain, and hence introduce another infinitesimal generator, which is denoted by $\mathcal{Q}^{\star}$. The modified chain is characterised by the presence of *one permanently active data call*, i.e. there is one active data call that never leaves the system, but shares in the available traffic channels as if it were a regular active data call. The behaviour of this permanent data call, i.e. the tagged call whose transfer time is to be determined, is identical to that of a regular data call, except for the fact that it cannot terminate within a given short time Δ which is considered in the proof of Lemma 2.1 below. Infinitesimal generator $\mathcal{Q}^{\star}$ is similar to $\mathcal{Q}$, but of smaller dimensions, since the rows and columns corresponding to all states $(s,d) \notin \mathbb{S}_{+}^{t+}$ are crossed out. In order to obtain $\mathcal{Q}^{\star}$, for all $(s,d) \in \mathbb{S}_+^{t+}$, the data call departure rates are modified (compared to the 'original' generator $\mathcal{Q}$) as follows:

$$\mathcal{Q}^\star\left((s,d);(s,d-1)\right) = \beta(s,d)\left(d_t(s,d)-1\right)\mu_{\text{data}},$$

and the diagonal elements of $\mathcal{Q}^\star$ are such that the entries of each row of $\mathcal{Q}^\star$ sum up to 0.

Furthermore, $\mathcal{B} \equiv diag(\beta(s,d),\ (s,d) \in \mathbb{S}_+^{t+})$ is the diagonal matrix of average data channel assignments, lexicographically ordered in (s,d). Note that, since $\beta(s,d) > 0$ for all $(s,d) \in \mathbb{S}_+^{t+}$, the diagonal matrix $\mathcal{B}$ is non-singular and thus $\mathcal{B}^{-1}$ exists.

We may now formulate the following lemma.

Lemma 2.1 *For $x \geq 0$, the vector of conditional expected transfer times $\widehat{\boldsymbol{\tau}}(x)$ satisfies the following differential equation and initial condition:*

$$\begin{aligned} \frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}(x) &= \mathcal{B}^{-1}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^\star\widehat{\boldsymbol{\tau}}(x), && (2.5)\\ \widehat{\boldsymbol{\tau}}(0) &= \mathbf{0}. && (2.6)\end{aligned}$$

Proof The lemma is proven by marginal analysis. Consider a time interval of length $\Delta > 0$, with Δ sufficiently small such that the tagged call cannot terminate within this time, hence $\Delta < x/\overline{\beta}(s,d)$ in state $(s,d) \in \mathbb{S}_+^{t+}$. Recall the definitions and properties of $\underline{\beta}(s,d)$ and $\overline{\beta}(s,d)$ from Section 2.4.2. Condition on all the possible events occurring in this interval, starting in state $(s,d) \in \mathbb{S}_+^{t+}$. For notational convenience and readability, the boundary constraints are not explicitly considered. Equations for the boundary can be derived analogously.

$$\begin{aligned} \widehat{\tau}_{s,d}(x) = \ & \Delta \\ & + \lambda_{\text{speech}}\,\Delta\,\widehat{\tau}_{s+1,d}(x - O(\Delta)) \\ & + s\mu_{\text{speech}}\,\Delta\,\widehat{\tau}_{s-1,d}(x - O(\Delta)) \\ & + \lambda_{\text{data}}\,\Delta\,\widehat{\tau}_{s,d+1}(x - O(\Delta)) \\ & + (\beta(s,d)\,d_t(s,d) - \underline{\beta}(s,d))\,\underline{p}(s,d)\,\mu_{\text{data}}\,\Delta\,\widehat{\tau}_{s,d-1}(x - O(\Delta)) \\ & + (\beta(s,d)\,d_t(s,d) - \overline{\beta}(s,d))\,\overline{p}(s,d)\,\mu_{\text{data}}\,\Delta\,\widehat{\tau}_{s,d-1}(x - O(\Delta)) \end{aligned}$$

$$\begin{aligned}
&+\underline{p}(s,d)\,(1-(\lambda_{\text{speech}}+s\,\mu_{\text{speech}}+\lambda_{\text{data}}+(\beta(s,d)\,d_t(s,d)+\\
&\qquad\qquad\qquad -\underline{\beta}(s,d))\mu_{\text{data}}\,)\Delta)\,\widehat{\tau}_{s,d}(x-\underline{\beta}(s,d)\Delta)\\
&+\overline{p}(s,d)\,(1-(\lambda_{\text{speech}}+s\,\mu_{\text{speech}}+\lambda_{\text{data}}+(\beta(s,d)\,d_t(s,d)+\\
&\qquad\qquad\qquad -\overline{\beta}(s,d))\,\mu_{\text{data}})\Delta)\,\widehat{\tau}_{s,d}(x-\overline{\beta}(s,d)\Delta)\\
&+o(\Delta),
\end{aligned}$$

where the Landau symbol $O(\Delta)$ ($o(\Delta)$) is standard notation for some unspecified function $F(\Delta)$ ($f(\Delta)$) having the property that $\lim_{\Delta\to 0} F(\Delta) = 0$ ($\lim_{\Delta\to 0} f(\Delta)/\Delta = 0$), i.e. $F(\Delta)$ ($f(\Delta)$) becomes negligibly small (compared to Δ) as $\Delta \to 0$, i.e. $F(\Delta)$ ($f(\Delta)$) becomes negligibly small (compared to Δ) as $\Delta \to 0$. The fifth and seventh line on the right hand side correspond with the case that the tagged data call is assigned $\underline{\beta}(s,d)$ channels, while in the sixth and the eighth line the tagged call is assigned one extra channel, i.e. $\overline{\beta}(s,d)$ channels.

Rearranging terms, and substituting (2.1), yields

$$\begin{aligned}
&\underline{p}(s,d)\,\underline{\beta}(s,d)\,\frac{\widehat{\tau}_{s,d}(x)-\widehat{\tau}_{s,d}(x-\underline{\beta}(s,d)\,\Delta)}{\underline{\beta}(s,d)\,\Delta}\\
&\quad+\overline{p}(s,d)\,\overline{\beta}(s,d)\,\frac{\widehat{\tau}_{s,d}(x)-\widehat{\tau}_{s,d}(x-\overline{\beta}(s,d)\,\Delta)}{\overline{\beta}(s,d)\,\Delta}\\
&=\ 1\\
&\qquad+\ \lambda_{\text{speech}}\,\widehat{\tau}_{s+1,d}(x-O(\Delta))\\
&\qquad+\ s\,\mu_{\text{speech}}\,\widehat{\tau}_{s-1,d}(x-O(\Delta))\\
&\qquad+\ \lambda_{\text{data}}\,\widehat{\tau}_{s,d+1}(x-O(\Delta))\\
&\qquad+\ \beta(s,d)\,(d_t(s,d)-1)\,\mu_{\text{data}}\,\widehat{\tau}_{s,d-1}(x-O(\Delta))\\
&\qquad+\ \underline{p}(s,d)\,(-\lambda_{\text{speech}}-s\,\mu_{\text{speech}}-\lambda_{\text{data}}-(\beta(s,d)\,d_t(s,d)+\\
&\qquad\qquad\qquad -\underline{\beta}(s,d))\,\mu_{\text{data}})\widehat{\tau}_{s,d}(x-\underline{\beta}(s,d)\,\Delta)\\
&\qquad+\ \overline{p}(s,d)\,(-\lambda_{\text{speech}}-s\,\mu_{\text{speech}}-\lambda_{\text{data}}-(\beta(s,d)\,d_t(s,d)+\\
&\qquad\qquad\qquad -\overline{\beta}(s,d))\,\mu_{\text{data}})\widehat{\tau}_{s,d}(x-\overline{\beta}(s,d)\,\Delta)\\
&\qquad+\ o(\Delta)/\Delta.
\end{aligned}$$

Letting $\Delta \downarrow 0$, and substituting (2.1) once again, gives

$$\begin{aligned}
&\beta(s,d)\,\frac{\partial \widehat{\tau}_{s,d}(x)}{\partial x}\\
&= \underline{p}(s,d)\,\underline{\beta}(s,d)\,\lim_{\Delta\downarrow 0}\left(\frac{\widehat{\tau}_{s,d}(x)-\widehat{\tau}_{s,d}(x-\underline{\beta}(s,d)\,\Delta)}{\underline{\beta}(s,d)\,\Delta}\right)\\
&\quad -\overline{p}(s,d)\,\overline{\beta}(s,d)\,\lim_{\Delta\downarrow 0}\left(\frac{\widehat{\tau}_{s,d}(x)-\widehat{\tau}_{s,d}(x-\overline{\beta}(s,d)\,\Delta)}{\overline{\beta}(s,d)\,\Delta}\right)\\
&= 1\\
&\quad +\lambda_{\text{speech}}\,\widehat{\tau}_{s+1,d}(x)\\
&\quad +s\,\mu_{\text{speech}}\,\widehat{\tau}_{s-1,d}(x)\\
&\quad +\lambda_{\text{data}}\,\widehat{\tau}_{s,d+1}(x)\\
&\quad +\beta(s,d)\,(d_t(s,d)-1)\,\mu_{\text{data}}\,\widehat{\tau}_{s,d-1}(x)\\
&\quad +\left(-\lambda_{\text{speech}}-s\,\mu_{\text{speech}}-\lambda_{\text{data}}-\beta(s,d)\,(d_t(s,d)-1)\,\mu_{\text{data}}\right)\,\widehat{\tau}_{s,d}(x).
\end{aligned}$$

Note that both limits are well-defined since $\underline{\beta}(s,d), \overline{\beta}(s,d) > 0$. This system of differential equations may equivalently be written in matrix notation,

$$\mathcal{B}\,\frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}(x) = \mathbf{1} + \mathcal{Q}^{\star}_{\text{data}}\widehat{\boldsymbol{\tau}}(x) \Leftrightarrow \frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}(x) = \mathcal{B}^{-1}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^{\star}_{\text{data}}\widehat{\boldsymbol{\tau}}(x).$$

To conclude the proof, the initial condition simply reflects the fact that the transfer time $\tau_{s,d}(0)$ of an ‘empty’ data call is zero, almost surely. □

Proposition 2.4 below presents the explicit expression of the conditional expected transfer time.

Proposition 2.4 *Let $\boldsymbol{\pi}^{\star} \equiv \left(\pi^{\star}_{s,d},\ (s,d) \in \mathbb{S}^{a+}_{+}\right)$ be the stationary probability distribution vector corresponding to the Markov chain with one permanently active data call, i.e. $\boldsymbol{\pi}^{\star}\mathcal{Q}^{\star} = \mathbf{0}$. Further, let $\boldsymbol{\gamma} = (\gamma_{s,d},\ (s,d) \in \mathbb{S}^{t+}_{+})$ be the unique solution to*

$$\begin{aligned}
\mathcal{Q}^{\star}\boldsymbol{\gamma} &= \frac{\mathcal{B}\mathbf{1}}{\boldsymbol{\pi}^{\star}\mathcal{B}\,\mathbf{1}} - \mathbf{1}, && (2.7)\\
\boldsymbol{\pi}^{\star}\mathcal{B}\,\boldsymbol{\gamma} &= 0. && (2.8)
\end{aligned}$$

Then the unique solution to the system of differential equations (2.5) is given by

$$\widehat{\boldsymbol{\tau}}(x) = \frac{x}{\boldsymbol{\pi}^\star \mathcal{B}\,\mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\right]\boldsymbol{\gamma}. \tag{2.9}$$

Proof In order to prove that the system of differential equations (2.5) with initial condition (2.6) has a *unique* solution, note that it is a system of the form $\frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}(x) = \mathbf{a}_o + \mathcal{A}\,\widehat{\boldsymbol{\tau}}(x) \equiv f(\widehat{\boldsymbol{\tau}}(x))$ where f is a linear function with continuous partial derivatives with respect to the entries of its argument vector. The existence and uniqueness of a solution $\widehat{\boldsymbol{\tau}}(x)$ for every initial vector, immediately follows from e.g. [51, Chapter 1, Section 8].

The existence of a vector $\boldsymbol{\gamma}$ that satisfies (2.7) and its uniqueness up to a translation along the vector $\mathbf{1}$, are guaranteed by results in Markov reward chain theory. Interpreting $\boldsymbol{\gamma}$ as the vector of relative rewards in a Markov reward chain governed by the infinitesimal generator $\mathcal{Q}^\star$ and with immediate reward vector $\frac{1}{\eta}\left(\frac{\mathcal{B}\mathbf{1}}{\boldsymbol{\pi}^\star\mathcal{B}\mathbf{1}} - \mathbf{1}\right)$ where η is the maximum rate of change in the Markov chain, and understanding that the long-term average rewards are zero, $\boldsymbol{\pi}^\star\left(\frac{\mathcal{B}\mathbf{1}}{\boldsymbol{\pi}^\star\mathcal{B}\mathbf{1}} - \mathbf{1}\right) = 1 - 1 = 0$, e.g. [213, Theorem 3.1, page 167] can be directly applied after uniformisation of the continuous-time Markov chain. In each state, the relative reward is equal to the long-run difference in accumulated rewards when starting in that state, relative to the rewards earned when starting in steady state. Note that indeed a translation of $\boldsymbol{\gamma}$ along the vector $\mathbf{1}$ does not alter the solution, since $\exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\mathbf{1} = \mathbf{1}$, which is readily verified using the Taylor expansion of $\exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}$. Hence in (2.7) a single degree of freedom exists in choosing $\boldsymbol{\gamma}$, which is used to normalise $\boldsymbol{\gamma}$ as in (2.8).

The proposition is then proven by substituting the claimed unique solution into the system of differential equations and verifying whether it indeed holds. With $\widehat{\boldsymbol{\tau}}(x)$ as given in (2.9),

$$\begin{aligned}
\frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}(x) &= \frac{1}{\boldsymbol{\pi}^\star\mathcal{B}\,\mathbf{1}}\mathbf{1} - \exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\mathcal{B}^{-1}\mathcal{Q}^\star\boldsymbol{\gamma} \\
&= \mathcal{B}^{-1}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^\star\left[\mathcal{I} - \exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\right]\boldsymbol{\gamma} \\
&= \mathcal{B}^{-1}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^\star\frac{x}{\boldsymbol{\pi}^\star\mathcal{B}\,\mathbf{1}}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^\star\left[\mathcal{I} - \exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\right]\boldsymbol{\gamma} \\
&= \mathcal{B}^{-1}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^\star\left[\frac{x}{\boldsymbol{\pi}^\star\mathcal{B}\,\mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\right]\boldsymbol{\gamma}\right] \\
&= \mathcal{B}^{-1}\mathbf{1} + \mathcal{B}^{-1}\mathcal{Q}^\star\widehat{\boldsymbol{\tau}}(x),
\end{aligned}$$

where (2.7) is substituted to obtain the second equality, and the third equality follows from adding $\mathcal{B}^{-1}\mathcal{Q}^{\star}\frac{x}{\boldsymbol{\pi}^{\star}\mathcal{B}\mathbf{1}}\mathbf{1}$, which is equal to zero, since $\mathcal{Q}^{\star}$ is an infinitesimal generator of a Markov chain, and hence $\mathcal{Q}^{\star}\mathbf{1} = \mathbf{0}$. To conclude the proof, observe that (2.9) satisfies the initial condition (2.6). □

Remark 2.2 The constant $\boldsymbol{\pi}^{\star}\mathcal{B}\,\mathbf{1} = \sum_{(s,d)\in\mathbb{S}_{+}^{t+}} \beta(s,d)\,\pi^{\star}(s,d)$ can be interpreted as the expected number of channels that are assigned to the permanently active data call, in the modified Markov chain, generated by $\mathcal{Q}^{\star}$. An equivalent expression follows from

$$\begin{aligned}
\boldsymbol{\pi}^{\star}\mathcal{B}\,\mathbf{1} &= \sum_{(s,d)\in\mathbb{S}_{+}^{t+}} d_t(s,d)\,\beta(s,d)\,\pi^{\star}(s,d) \\
&\qquad -\mu_{\text{data}}^{-1} \sum_{(s,d)\in\mathbb{S}_{+}^{t+}} (d_t(s,d)-1)\,\beta(s,d)\,\mu_{\text{data}}\,\pi^{\star}(s,d) \\
&= \sum_{(s,d)\in\mathbb{S}_{+}^{t+}} d_t(s,d)\,\beta(s,d)\,\pi^{\star}(s,d) \\
&\qquad -\mu_{\text{data}}^{-1} \sum_{(s,d)\in\mathbb{S}_{+}^{t+}} \lambda_{\text{data}}\,1\{(s,d+1)\in\mathbb{S}_{+}^{t+}\}\,\pi^{\star}(s,d) \\
&\equiv \mathbf{C}_{\text{data}}^{\star} - \rho_{\text{data}}(1-\mathbf{P}_{\text{data}}^{\star})
\end{aligned}$$

where $\mathbf{C}_{\text{data}}^{\star}$ is the average number of channels used for data transfer, and $\mathbf{P}_{\text{data}}^{\star}$ is the blocking probability of a newly arriving data call, both in the modified Markov chain. In the derivation above, the second equality sign is due to the fact that in steady state the average number of data calls leaving the system per time unit must equal the average number of data calls entering the system per time unit. Note that since $\rho_{\text{data}}(1-\mathbf{P}_{\text{data}}^{\star})$ is the average number of channels used by non-permanent data calls (as is clear from the above derivation), $\mathbf{C}_{\text{data}}^{\star} - \rho_{\text{data}}(1-\mathbf{P}_{\text{data}}^{\star})$ is indeed equal to the expected number of channels assigned to the permanent data call.

Remark 2.3 Although the proof of Proposition 2.4 may seem to indicate that instead of the constant $\boldsymbol{\pi}^{\star}\mathcal{B}\,\mathbf{1}$ basically any value could have been used, we stress that $\boldsymbol{\pi}^{\star}\mathcal{B}\mathbf{1}$ is indeed the only constant that allows (2.7) to be solved for $\boldsymbol{\gamma}$, as is easily demonstrated by premultiplication of (2.7) by $\boldsymbol{\pi}^{\star}\mathcal{B}$.

Corollary 2.2 below derives an asymptotic result, establishing an additional fairness property of the investigated class of channel sharing policies in that the transfer time of a data call is approximately linear in the data call size.

Corollary 2.2 *The following asymptotic result immediately follows from (2.9):*

$$\lim_{x\to\infty}\left\{\widehat{\boldsymbol{\tau}}(x)-\frac{x}{\boldsymbol{\pi}^\star\mathcal{B}\,\mathbf{1}}\mathbf{1}\right\}=\boldsymbol{\gamma}.$$

Proof To see this, note that $\mathcal{B}^{-1}\mathcal{Q}^\star$ is the infinitesimal generator of an irreducible finite state space Markov chain, with equilibrium distribution vector $\frac{\boldsymbol{\pi}^\star\mathcal{B}}{\boldsymbol{\pi}^\star\mathcal{B}\mathbf{1}}$. Hence $\lim_{x\to\infty}\exp\left\{x\mathcal{B}^{-1}\ \ \mathcal{Q}^\star\right\}=\mathbf{1}\frac{\boldsymbol{\pi}^\star\mathcal{B}}{\boldsymbol{\pi}^\star\mathcal{B}\mathbf{1}}$, the matrix with each row equal to the equilibrium distribution vector, and $\lim_{x\to\infty}\left(\mathcal{I}-\exp\ \left\{x\mathcal{B}^{-1}\mathcal{Q}^\star\right\}\right)\boldsymbol{\gamma}=\boldsymbol{\gamma}-\mathbf{1}\frac{\boldsymbol{\pi}^\star\mathcal{B}}{\boldsymbol{\pi}^\star\mathcal{B}\mathbf{1}}\boldsymbol{\gamma}=\boldsymbol{\gamma}$ using (2.8). Note that this simplest form of the asymptotic result follows from normalising $\boldsymbol{\gamma}$ as is done in (2.8). □

Remark 2.4 The asymptotic result presented in Corollary 2.2 is readily supported by the following intuitive argument. Consider a call of size x, expressed as the transfer time in seconds given the exclusive use of one traffic channel. As $x\to\infty$ the average number of assigned channels over the call's lifetime becomes more and more independent of the system state at the call's arrival and the precise evolution trace of all other (speech or data) calls. In fact, in the limit the average number of assigned channels is precisely equal to $\boldsymbol{\pi}^\star\mathcal{B}\,\mathbf{1}$, and hence the expected transfer time is equal to the deterministic call size divided by the average number of assigned traffic channels. We note hereby that the significance of the constant $\boldsymbol{\gamma}$ becomes negligible for $x\to\infty$.

2.7. CHANNEL SHARING POLICIES

In this section we propose four distinct channel sharing policies for speech and data calls that fit within the considered class. The first policy, called SEGREGATION, is a *static* policy, in that speech and data calls are served with two completely separate channel pools. The other policies, FIXED, SHARE and SHARE-RESERVE allow *dynamic* sharing of the traffic channels, to various extents. For each policy the basic functions $s_{\max}(s,d)$, $d_{\max}(s)$ and $\beta(s,d)$ are specified as well as $s_{\max}$ and $d_{\max}$, while we recall that $d_t(s,d)$ and $d_a(s,d)$ are implicitly defined by these basic functions.

2.7.1. SEGREGATION POLICY

Under the SEGREGATION policy, the speech and data services are completely segregated, in that the cell capacity of C_{total} traffic channels is split into two disjoint pools with $C_{\text{speech}} \geq 1$ and $C_{\text{data}} \equiv C_{\text{total}} - C_{\text{speech}} \geq b$ channels for speech and data calls, respectively. Such a policy may be preferred in view of the low implementational complexity. Since there is no interaction between the two service types, the performance analysis can be done separately. Still, in order to demonstrate that the SEGREGATION policy falls within the studied class of channel sharing policies, the three characteristic functions will be specified.

For *speech* calls, the resulting model is simply an $M/M/C_{\text{speech}}/C_{\text{speech}}$ Erlang loss model with speech traffic load ρ_{speech}. Note that $s_{\max} \equiv s_{\max}(0,0) = s_{\max}(s,d) = C_{\text{speech}}$. The speech call blocking probability is given by the well-known Erlang loss formula, e.g. [213]. *Data* calls request a transfer capacity of $\beta^{\max}_{\text{HSCSD}}$ channels, but will settle for any capacity between $\beta^{\min}$ and $\beta^{\max}_{\text{HSCSD}}$. During a data call, the channel assignment is dynamically adapted to either utilise freed capacity or to support newly admitted data calls. Hence the average number of channels assigned to an active data call in system state (s,d), is given by $\beta(s,d) \equiv \min\left\{\beta^{\max}_{\text{HSCSD}}, \frac{C_{\text{data}}}{d_t(s,d)}\right\}$ if $d_t(s,d) > 0$. The maximum number of data calls in the segregated system is given by $d_{\max} \equiv d_{\max}(0) = d_{\max}(s) = \left\lfloor \frac{C_{\text{data}}}{\beta^{\min}} \right\rfloor + Q_a$, which functions as the Call Admission Control threshold. Observe that all these functions do not depend on s.

In this fully segregated model, the speech and data services can be evaluated separately, but all performance measures obtained are identical to those that would be found if the segregated models were evaluated simultaneously as one model.

2.7.2. FIXED POLICY

Under the FIXED policy, data calls request a fixed capacity of $\beta_{\text{FIXED}} \in \{\beta^{\min}, \cdots, \beta^{\max}_{\text{HSCSD}}\}$ traffic channels. Speech service is protected from the potentially demanding data calls, by reserving C_{speech} channels for speech calls only. The remaining $C_{\text{total}} - C_{\text{speech}}$ channels are shared between speech and data calls, without any priorities or service preemption.

Speech calls are admitted if at least 1 channel is available, i.e. if $s < s_{\max}(s,d) \equiv C_{\text{total}} - \beta_{\text{FIXED}}\, d_t(s,d)$. The maximum number of speech calls in the system is given by $s_{\max} \equiv s_{\max}(0,0) = C_{\text{total}}$. A *data* call is admitted if it can either start service immediately, i.e. β_{FIXED} free channels can be found among the $C_{\text{total}} - C_{\text{speech}}$ shared channels, or if the access queue is not full. Mathematically, this condition for data call admission is formulated as follows: a data call is admitted if $d < d_{\max}(s) \equiv \left\lfloor \frac{C_{\text{total}} - \max\{C_{\text{speech}}, s\}}{\beta_{\text{FIXED}}} \right\rfloor + Q_a$. Once activated, data calls hold on to the assigned $\beta(s,d) \equiv \beta_{\text{FIXED}}$ channels until call termination, continuously transmitting at a fixed bit rate. The maximum number of data calls in the system is given by $d_{\max} \equiv d_{\max}(0) = \left\lfloor \frac{C_{\text{total}} - C_{\text{speech}}}{\beta_{\text{FIXED}}} \right\rfloor + Q_a$. Note that since $\beta(s,d) \equiv \beta_{\text{FIXED}}$ indicates a fixed transfer rate in each state $(s,d) \in \mathbb{S}_{+}^{t+}$, $\boldsymbol{\gamma} = \mathbf{0}$ immediately follows from (2.7) and (2.8), so that (2.9) yields $\widehat{\boldsymbol{\tau}}(x) = \beta_{\text{FIXED}}^{-1}\, x\, \mathbf{1}$, as expected.

The FIXED channel sharing policy is different from the three other proposed policies in the sense that the data calls are not elastic, i.e. the policy does not allow the data calls to dynamically capture or release traffic channels, in order to enhance service quality and channel utilisation, or support newly arriving (speech or data) calls. Hence the associated performance model is a rather basic loss model with a limited access queue. The FIXED policy is included in the performance comparison for reference purposes.

2.7.3. SHARE POLICY

Under the SHARE policy, data calls request a transfer capacity of $\beta_{\text{HSCSD}}^{\max}$ channels, but will settle for any capacity between $\beta^{\min}$ and $\beta_{\text{HSCSD}}^{\max}$. Data calls maximally utilise all available channels, with channel assignments that are dynamically adapted to either utilise freed capacity or to support newly admitted (speech or data) calls. The cell capacity C_{total} is fully shared between speech and data calls, whereby data calls are always forced to give up excess capacity, i.e. capacity above $\beta^{\min}$, when needed. An important distinction between the SHARE policy and the other proposed policies, is that here no parameters need to be set by the network operator.

Speech calls are admitted if at least 1 channel is, or can be made, available, i.e. if $s < s_{\max}(s,d) \equiv C_{\text{total}} - \beta^{\min} d_t(s,d)$. Note that under this policy, if $\beta^{\min} = 1$ speech calls can be admitted only if no data calls are queued, while for $\beta^{\min} > 1$ it is possible for a speech call to be admitted, even if one or more data calls are queued. The maximum number of speech calls in the system is given by $s_{\max} \equiv s_{\max}(0,0) = C_{\text{total}}$. A *data* call is admitted if at least $\beta^{\min}$ channels are, or can be made, available, or if the access queue is not full. This condition can be mathematically formulated as follows: $d < d_{\max}(s) \equiv \left\lfloor \frac{C_{\text{total}}-s}{\beta^{\min}} \right\rfloor + Q_a$. Once activated, each active data call will receive $\min\left\{\beta^{\max}_{\text{HSCSD}}, \left\lfloor \frac{C_{\text{total}}-s}{d_t(s,d)} \right\rfloor\right\} \geq \beta^{\min}$ channels, while the remaining channels are randomly distributed, not exceeding the technical constraint $\beta^{\max}_{\text{HSCSD}}$. Thus on average each active data call is given $\beta(s,d) \equiv \min\left\{\beta^{\max}_{\text{HSCSD}}, \frac{C_{\text{total}}-s}{d_t(s,d)}\right\}$ channels if $d_t(s,d) > 0$. The maximum number of data calls in the system is given by $d_{\max} \equiv d_{\max}(0) = \left\lfloor \frac{C_{\text{total}}}{\beta^{\min}} \right\rfloor + Q_a$.

In a typical scenario with e.g. $\beta^{\min} = 1$, we note that as the system becomes overloaded with data traffic, i.e. $\lambda_{\text{data}} \to \infty$, the speech call blocking probability approaches 100% due to the fact that a freed channel will always be claimed immediately by a queued data call (provided that $Q_a > 0$). This suggests the need to protect the speech service, which is precisely the aim of the SHARE-RESERVE policy.

2.7.4. SHARE-RESERVE POLICY

The SHARE-RESERVE policy is very similar to the SHARE policy, except that now speech calls are strictly prioritised over data calls, in the sense that data calls are forced to release any assigned channels in the shared channel territory, in support of a newly admitted speech call. In order to prevent speech calls from crowding out data calls, C_{data} channels are reserved for the data service only, while the remaining $C_{\text{speech}} \equiv C_{\text{total}} - C_{\text{data}}$ channels are shared.

Speech calls are admitted if at least 1 channel is, or can be made, available, i.e. if $s < s_{\max}(s,d) \equiv C_{\text{total}} - C_{\text{data}}$. The maximum number of speech calls in the system is given by $s_{\max} \equiv s_{\max}(0,0) = C_{\text{total}} - C_{\text{data}}$. A *data* call is admitted if a minimum capacity of $\beta^{\min}$ channels can be guaranteed to it within the reserved territory of C_{data} channels, or if the access queue is not full. Mathematically formulated, the condition is as follows: $d < d_{\max}(s) \equiv \left\lfloor \frac{C_{\text{data}}}{\beta^{\min}} \right\rfloor + Q_a$. Once activated, data calls share the available channels fairly, precisely as described for the SHARE policy. Hence $\beta(s,d) \equiv \min\left\{\beta^{\max}_{\text{HSCSD}}, \frac{C_{\text{total}}-s}{d_t(s,d)}\right\} \geq \beta^{\min}$, for $d_t(s,d) > 0$. Note that no more than $\left\lfloor \frac{C_{\text{data}}}{\beta^{\min}} \right\rfloor$ data calls can be active, even if there are no speech calls in the system. The reason for this is that no more data calls can be *guaranteed* a minimum assignment of $\beta^{\min}$ channels, if a large number of speech calls were to arrive and claim all $C_{\text{total}} - C_{\text{data}}$ shared channels (recall that an active data call cannot be pushed back into the access queue). As a consequence, the maximum number of data calls in the system is given by $d_{\max} \equiv d_{\max}(0) = \left\lfloor \frac{C_{\text{data}}}{\beta^{\min}} \right\rfloor + Q_{\text{data}}$.

2.7.5. OVERVIEW OF CHANNEL SHARING POLICIES

We conclude this section with an overview of the four presented channel sharing policies. Figure 2.6 graphically summarises how the cell capacity C_{total} can be assigned to the different services, while Table 2.1 lists all functions required for the performance analysis. The functions $\beta(s,d)$ in this table are defined only if $d_t(s,d) > 0$.

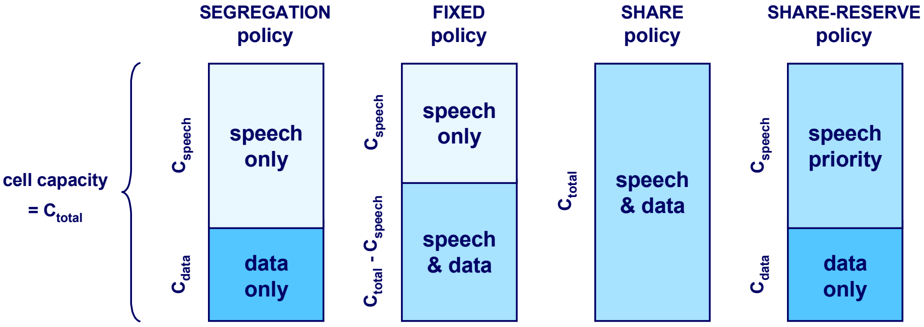


Figure 2.6 Overview of channel sharing policies: the SEGREGATION, FIXED, SHARE and SHARE-RESERVE policies feature different channel pool partitionings and priority schemes.

Table 2.1 Overview of channel sharing policies: specification of the channel assignment and service-specific CAC functions associated with the SEGREGATION, FIXED, SHARE and SHARE-RESERVE policies.

	$\beta(s,d)$	$s_{\max}(s,d)$	$d_{\max}(s) - Q_a$
SEGREGATION	$\min\left\{\beta^{\max}_{\text{HSCSD}}, \frac{C_{\text{data}}}{d_t(s,d)}\right\}$	C_{speech}	$\left\lfloor \frac{C_{\text{data}}}{\beta^{\min}} \right\rfloor$
FIXED	β_{FIXED}	$C_{\text{total}} - \beta_{\text{FIXED}} d_t(s,d)$	$\left\lfloor \frac{C_{\text{total}} - \max\{C_{\text{speech}}, s\}}{\beta_{\text{FIXED}}} \right\rfloor$
SHARE	$\min\left\{\beta^{\max}_{\text{HSCSD}}, \frac{C_{\text{total}} - s}{d_t(s,d)}\right\}$	$C_{\text{total}} - \beta^{\min}\, d_t(s,d)$	$\left\lfloor \frac{C_{\text{total}} - s}{\beta^{\min}} \right\rfloor$
SHARE-RESERVE	$\min\left\{\beta^{\max}_{\text{HSCSD}}, \frac{C_{\text{total}} - s}{d_t(s,d)}\right\}$	$C_{\text{total}} - C_{\text{data}}$	$\left\lfloor \frac{C_{\text{data}}}{\beta^{\min}} \right\rfloor$

2.8. NUMERICAL RESULTS

This section presents an extensive numerical study. As it would take up too much space to study the effect of all model parameters, some parameters are prefixed at a realistic level while the remaining parameters are varied within a realistic range around their default value and their impact on the relevant performance measures is investigated. Table 2.2 below gives an overview of all model parameters and indicates either their prefixed value or their default values and the range of considered values around this default value.

Table 2.2 Numerical results: the default settings and investigated parameter ranges of the system and traffic parameters.

PARAMETER	PREFIXED VALUE		PARAMETER	DEFAULT VALUE		RANGE
r_{data}	14.4	kbits/s	C_{total}	21	channels	$\{7, 14, 21, 28\}$
$\beta^{\min}$	1	channel	C_{data}	6	channels	$\{2, 4, 6, 8\}$
$\beta^{\max}_{\text{HSCSD}}$	4	channels	C_{speech}	$C_{\text{total}} - C_{\text{data}}$	channels	–
β_{FIXED}	2	channels	Q_a	5	calls	$\{0, \cdots, 20\}$
μ_{data}	0.0450	calls/s	ρ_{data}	$0.5 \cdot \rho_{\text{speech}}$	Erlang	$(0, 1] \cdot \rho_{\text{speech}}$
μ_{speech}	0.0200	calls/s	λ_{data}	$\mu_{\text{data}} \cdot \rho_{\text{data}}$	calls/s	–
			ρ_{speech}	$0.6113 \cdot C_{\text{total}}$	Erlang	–
			λ_{speech}	$\mu_{\text{speech}} \cdot \rho_{\text{speech}}$	calls/s	–

Regarding the prefixed parameters, r_{data} is based on the latest channel coding scheme for a full rate data traffic channel, $\beta^{\min}$ and $\beta^{\max}_{\text{HSCSD}}$ correspond to the expected multislot capabilities of an HSCSD terminal, and μ_{data} and μ_{speech} are set to correspond with an average e-mail size (320 kbits: $\mu^{-1}_{\text{data}} = 320/14.4$ seconds, assuming the latest 14.4 kbits/s GSM/HSCSD channel rate) and an average speech call holding time in a cell ($\mu^{-1}_{\text{speech}} = 50$ seconds) [101, 219]. Note that parameters C_{data} and C_{speech} are not required for all channel sharing policies (see Table 2.1). The speech call arrival rate λ_{speech} is chosen such that for a cell with 3 frequencies (allowing $C_{\text{total}} = 21$ traffic channels, given an assumed requirement of 3 control signalling channels) the speech call blocking probability is 1% provided that all C_{total} channels are available for speech transfer (for 3 frequencies: $\rho_{\text{speech}} = 12.837$ Erlang). For those cases with fewer or more frequencies, the speech traffic load is linearly adjusted as indicated in Table 2.2. Finally, the data traffic load ρ_{data} is varied between 0 Erlang and ρ_{speech}. Since μ_{data} is fixed, λ_{data} is adjusted to obtain the desired data traffic load.

In the remainder of this section, a number of numerical experiments is executed in order to obtain insight into the effect of the variable parameters on the performance measures. Regarding the applied units, it is noted that the access, transfer and sojourn times are expressed in *seconds*, while the data traffic load in the charts is expressed as a fraction of the speech traffic load (in *Erlang*), and the data call size is expressed in units of r_{data} kbits.

2.8.1. COMPARISON OF CHANNEL SHARING POLICIES

The proposed channel sharing policies are compared with default settings for all model parameters except for the data traffic load ρ_{data} which is varied between 0 Erlang and $\rho_{\text{speech}} = 12.837$ Erlang. As Figure 2.7 (left) shows, under low data traffic loads the channel utilisation is optimal under the FIXED and SHARE policies, since they do not reserve any capacity strictly for data transfers. As the data traffic load grows, however, only those policies that are closest to being truly work-conserving (SHARE and SHARE-RESERVE), are able to establish a significant channel utilisation, since under these policies data calls can occupy up to four otherwise idle (reserved for speech calls) traffic channels.

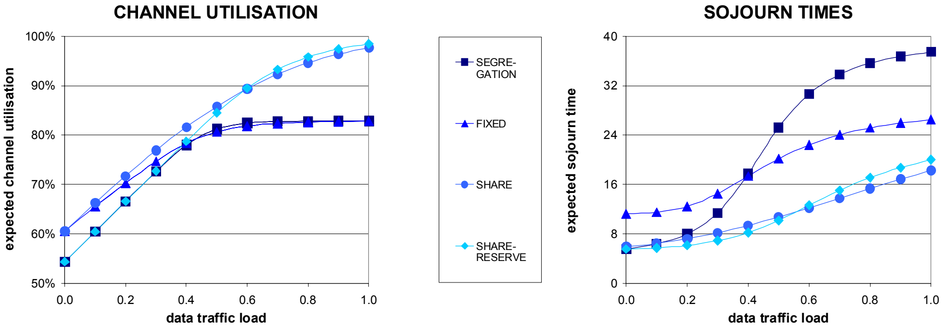


Figure 2.7 Comparison of channel sharing policies: expected channel utilisation versus data traffic load (left) and expected sojourn times versus data traffic load (right).

Under the same parameter settings, Figure 2.8 presents the speech (left) and data (right) call blocking probability as a function of the data traffic load for all four policies. This figure reveals the primary disadvantage of the SHARE policy in the sense that it cannot protect the speech service from being crowded out by the data traffic. A low data call blocking probability along with a rapidly increasing speech call blocking probability clearly indicates this. In contrast, under low data traffic loads the SHARE policy is optimal. Note from the channel utilisation and blocking probabilities that the performance of the FIXED policy converges to that of the SEGREGATION

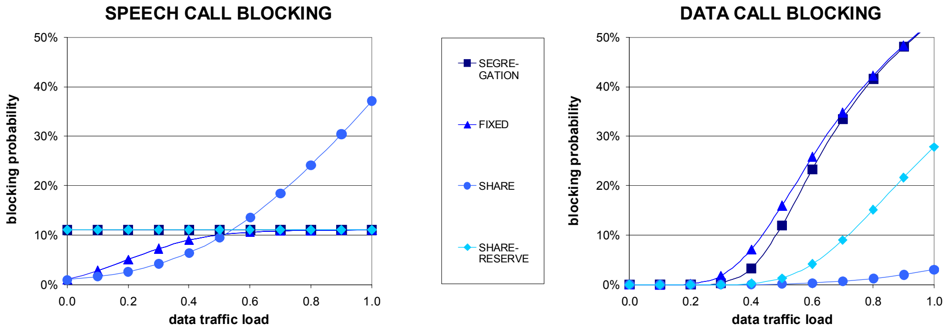


Figure 2.8 Comparison of channel sharing policies: speech (left) and data (right) call blocking probability versus data traffic load.

policy for increasing data traffic loads, because the data calls will fully occupy all channels available for data transfer. Since the SEGREGATION policy assigns only a single channel to each data call under high data traffic loads, the number of active data calls is twice as high while the transfer is twice as slow compared to the fixed assignment of two channels under the FIXED policy. The convergence of the channel utilisation and blocking probabilities proves that the two opposite effects cancel out.

Regarding the corresponding expected sojourn times, observe from Figure 2.7 (right) that for low data traffic loads, the FIXED policy is strictly outperformed by the other policies that allow assignments of more than β_{FIXED} traffic channels. As ρ_{data} grows, the expected sojourn times grow under each policy, including the FIXED policy due to an increasing access time component. The SEGREGATION policy, not allowing data calls to utilise idle speech channels, suffers from this restriction most notably under high data traffic loads, as indicated by the long expected sojourn times. The SHARE and SHARE-RESERVE policies yield very similar Quality of Service curves.

Table 2.3 provides an overview of the performance of the investigated channel sharing policies regarding the principal performance measures. Separately for low ($\rho_{\text{data}} \approx 0$) and high ($\rho_{\text{data}} \gg 0$) data traffic loads, a policy scores a '$-$', '0' or '+' reflecting the relative performance with respect to the other policies. Evidently, none of the policies strictly outperforms the alternatives with respect to all performance measures, which prohibits a trivial policy selection. We argue that both the work-conserving SHARE and SHARE-RESERVE policies are preferred over the SEGREGATION and the FIXED policies, primarily because they allow statistical multiplexing of speech and data traffic and thus generally establish a high channel utilisation and, correspondingly, low sojourn times and blocking probabilities. In the initial phase of a light data traffic load, it is recommended to implement the SHARE policy, as any channel reservation would only raise the speech call blocking probability. However, when the data traffic load grows from light to moderate or heavy, it appears best to deploy the SHARE-RESERVE policy, as a mobile network operator is likely to be very hesitant about affecting its speech client base when operating in the data market. The rationale for this is that the policy best utilises the elasticity and relative delay tolerance of the data calls, while protecting the speech users by posing an acceptable upper bound on the speech call blocking probability, which is independent from the data traffic load. Since it is most robust against a data traffic load increase, we select the SHARE-RESERVE policy for further study in the remainder of our numerical investigation. In practice, we suggest that a desired trade-off between the Grade and Quality Of Service measures is established by making the reservation level adaptive to the traffic load.

Table 2.3 Comparison of channel sharing policies.

	$\mathbf{U}$		$\mathbf{P}_{\text{speech}}$		$\mathbf{P}_{\text{data}}$		$\mathbf{T}_{\text{data}}$	
ρ_{data}	≈ 0	$\gg 0$	≈ 0	$\gg 0$	≈ 0	$\gg 0$	≈ 0	$\gg 0$
SEGREGATION	0	−	−	+	+	−	+	−
FIXED	+	−	+	+	+	−	−	0
SHARE	+	+	+	−	+	+	+	+
SHARE-RESERVE	0	+	−	+	+	0	+	+

2.8.2. PERFORMANCE EFFECTS OF C_{data}

This section focuses on the trade-off between the different performance measures as we reserve fewer or more traffic channels for data transfer. The data traffic load ρ_{data} is varied between 0 Erlang and $\rho_{\text{speech}} = 12.837$ Erlang, C_{data} is taken from $\{2, 4, 6\}$ while all other model parameters are set to their default values. As Figure 2.9 (left) shows, C_{data} must be adapted to the data traffic load if an operator wishes to maximise channel utilisation, in accordance with the upper envelope of the utilisation curves. As the data traffic load increases, C_{data} should be regularly incremented in order to keep the channel utilisation maximal. The trade-off is that aiming for optimal channel utilisation may imply unacceptable speech call blocking probabilities under high data traffic loads (see Figure 2.10 (left)).

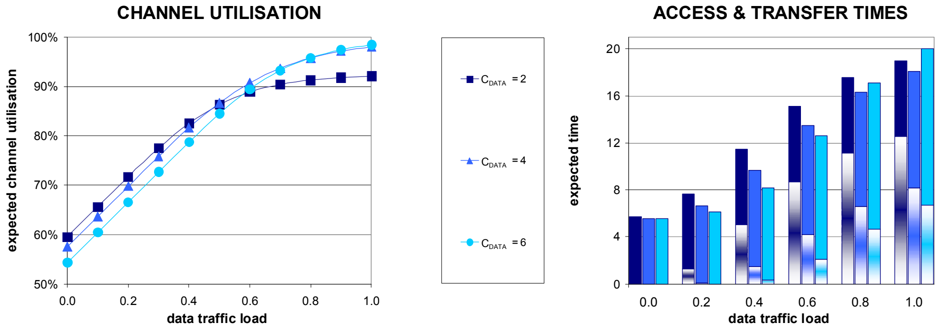


Figure 2.9 Performance effects of C_{data}: channel utilisation (left), expected access, transfer and sojourn times (right) versus data traffic load.

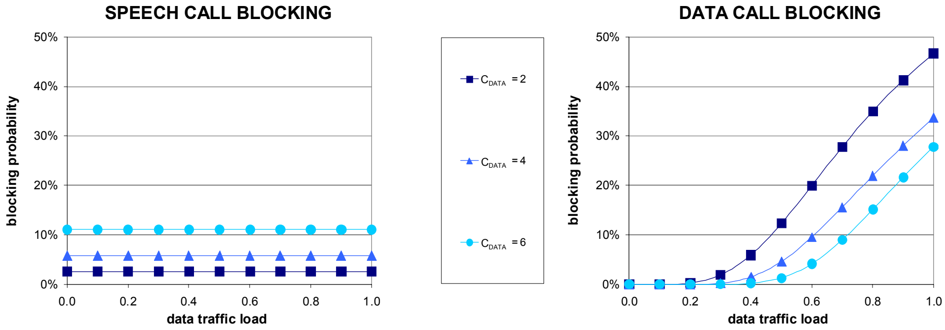


Figure 2.10 Performance effects of C_{data}: speech (left) and data (right) call blocking probability versus data traffic load.

Figure 2.9 (right) illustrates the effect that C_{data} has on the data calls' expected access, transfer and sojourn times. The height of each vertical bar reflects the expected sojourn time, consisting of a access time (bottom segment) and a transfer time (top segment) component. For very light data traffic loads, in particular for $\rho_{\text{data}} \downarrow 0$, delay values are plotted at the '$\rho_{\text{data}} = 0.0$' mark on the horizontal axis. Aside from the unsurprising results that the expected access, transfer and hence also the sojourn times increase with ρ_{data}, while the access time becomes more dominant with an increase in ρ_{data}, the figure also illustrates that an increase in C_{data} does *not* necessarily imply an enhancement of the delivered QOS for admitted data calls. The reason for this effect is that the data call blocking probability decreases with an increase in C_{data} (see Figure 2.10 (right)), so that the *admitted* data calls have more competition for radio resources.

Suppose the maximum speech and data call blocking probabilities an operator allows in its network are 5% and 10%. Then for the given setting, the presented performance results enable us to conclude that the optimal number of dedicated data traffic channels is $C_{\mathrm{data}} = 2$ for $\rho_{\mathrm{data}} \leq 0.45 \cdot 12.837$ Erlang and $C_{\mathrm{data}} = 4$ for $0.45 \cdot 12.837 < \rho_{\mathrm{data}} \leq 0.6 \cdot 12.837$ Erlang, while for greater data traffic loads, there is insufficient cell capacity to meet the blocking requirements. Note that under the proposed channel reservation strategy, the channel utilisation is maximal, while the expected data call sojourn times are still below 15 seconds. Viewing the problem from a slightly different angle, one can determine the maximum value of C_{data} such that the speech call blocking probability remains below a prespecified value, e.g. 5%, and subsequently determine the data call blocking probability and the expected sojourn times as demonstrated. If these performance measures for the data services are not satisfactory, the operator must increase the cell capacity, e.g. by assigning an additional frequency to it.

2.8.3. PERFORMANCE EFFECTS OF Q_a

We now investigate the performance effects of increasing the size of the access queue (Q_a) for the SHARE-RESERVE policy. Default settings are used for all model parameters except for the data traffic load ρ_{data} which is varied between 0 Erlang and $\rho_{\mathrm{speech}} = 12.837$ Erlang and Q_a which is taken from $\{0, 5, 10\}$. First, note that under the considered policy the speech call blocking probability is obviously independent of ρ_{data} and Q_a, as illustrated by Figure 2.12 (left). In contrast, the data call blocking probability (see Figure 2.12 (right)) increases with ρ_{data} and decreases with Q_a. For the given range of ρ_{data} the carried data traffic load $\rho_{\mathrm{data}}(1 - \mathbf{P}_{\mathrm{data}})$ still increases with the offered data traffic load ρ_{data}, which explains the channel utilisation, increasing to 100% (see Figure 2.11 (left)) as the data call blocking probability approaches 1 to stabilise the carried data load, independently of any further increase in ρ_{data}. Lastly, the lower data call blocking probability induced by a larger access queue marginally improves the channel utilisation.

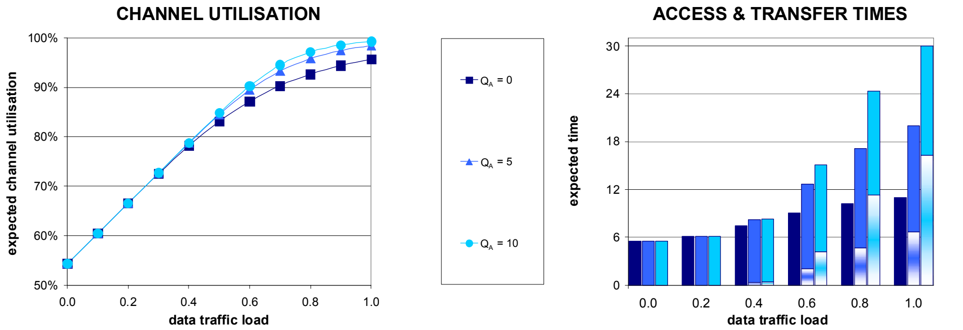


Figure 2.11 Performance effects of Q_a: channel utilisation (left), expected access, transfer and sojourn times (right) versus data traffic load.

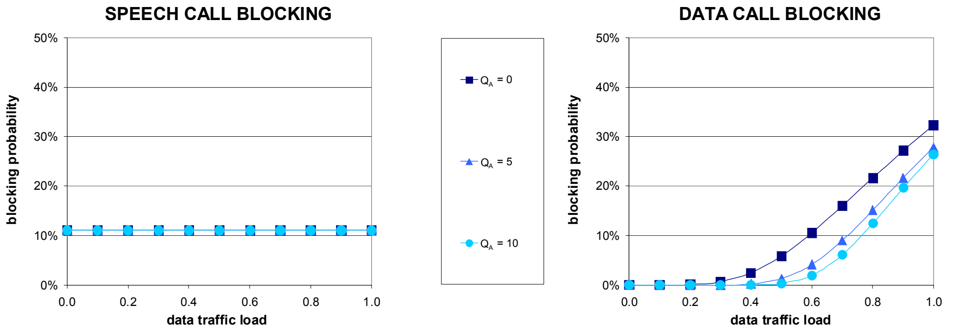


Figure 2.12 Performance effects of Q_a: speech (left) and data (right) call blocking probability versus data traffic load.

For $\rho_{\text{data}} \in \{0.0, 0.2, 0.4, 0.8, 1.0\} \cdot \rho_{\text{speech}}$ Figure 2.11 (right) presents the expected sojourn times and the corresponding break-down in access (bottom segment) and transfer (top segment) times. Obviously, there is no access time if there is no access queue ($Q_a = 0$). Once again, the delay values plotted at the '$\rho_{\text{data}} = 0.0$' mark on the horizontal axis must be interpreted as the limit values as $\rho_{\text{data}} \downarrow 0$. Under low data traffic loads, the sojourn time is dominated by the transfer time which is bounded from below by 5.56 seconds, corresponding with a continuous assignment of $\beta_{\text{HSCSD}}^{\max}$ traffic channels to each data call. As ρ_{data} increases, both the expected access and the transfer times go up, until the data traffic load becomes so high that each data call is served with no more than $\beta^{\min} = 1$ dedicated traffic channel *plus* a fair share of those shared channels that do not carry speech calls. For data traffic loads beyond this value, the expected transfer time remains constant at $C_{\text{data}} \cdot \left(\frac{320}{14.40\,(C_{\text{total}} - \mathbf{N}_{\text{speech}})}\right) \approx 13.91$ seconds with $\mathbf{N}_{\text{speech}} \approx 11.41$, $C_{\text{total}} - \mathbf{N}_{\text{speech}}$ the expected number of channels available for data transfer, and C_{data} the number of active data calls (recall that $\beta^{\min} = 1$). The expected access times continue to increase, converging to $Q_a \cdot \left(\frac{320}{14.40\,(C_{\text{total}} - \mathbf{N}_{\text{speech}})}\right)$, since in a cell overloaded with data calls, an admitted data call always takes the last position in the access queue and hence must wait until Q_a data calls finish their transfer. For $Q_a = 5$ and 10, the corresponding upper bounds on the expected access times are 11.59 and 23.18 seconds. Recalling that the expected transfer time was calculated to converge to 13.91, this illustrates that it depends on the access queue size whether the expected sojourn time will ever be dominated by the expected access time. For $Q_a = 10$ the access time is already dominant at $\rho_{\text{data}} = 12.837$ (see figure), while for $Q_a = 5$ the access time will never dominate.

2.8.4. ACCESS TIMES AND OPTIMISING Q_a

In the previous section it was illustrated that the expected access time was increasing in both the data traffic load ρ_{data} and the access queue size Q_a. In this section we focus on the effect of the access queue size on the expected access time and the data call blocking probability, and indicate how a network operator can choose the optimal value of its access queue size. A cell capacity C_{total} of $7, 14, 21$ or 28 traffic channels is considered with $C_{\text{data}} = 2, 4, 6$ or 8, respectively. The access queue size is varied from 0 to 20. All other model parameters are set to the default values. Recall that this implies proportionality of the speech and data traffic loads with respect to the cell capacity (see Table 2.2).

Figure 2.13 shows the speech (left) and data (right) call blocking probabilities. Naturally, the speech call blocking probability is unaffected by the variation in access queue size, while its dependency on the cell capacity is readily determined using the Erlang loss formula. The data call blocking probability decreases with the access queue size, while the expected access time (see Figure 2.14 (left)) converges to a constant as the access queue becomes so large that its last positions are virtually never taken (≈ 0% data call blocking probability). Regarding the effect of the cell capacity, the presented numerical results support the well-known result that the benefits of statistical multiplexing become greater as the capacity increases. This can be seen from the fact that although the offered load is assumed proportional to the cell capacity, cells with higher capacity are strictly better off with respect to both performance measures displayed.

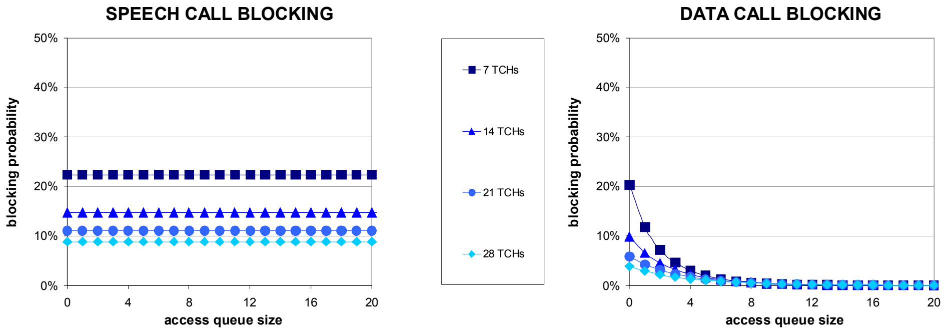


Figure 2.13 Access times and optimising Q_a: speech (left) and data (right) call blocking probability versus data traffic load.

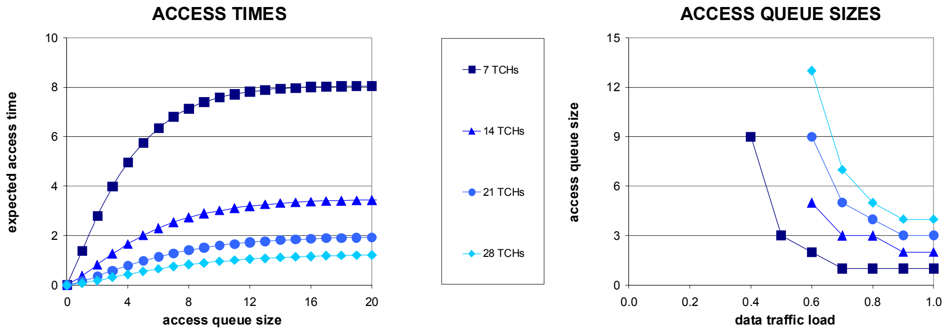


Figure 2.14 Access times and optimising Q_a: expected access times versus queue size (left) and maximum allowable queue sizes versus data traffic load (right).

Recall from Section 2.4 that the mere purpose of implementing an access queue for data call requests is to postpone call blocking momentarily in the hope that resources are freed to serve the call. The amount of additional set-up delay that can be allowed is limited. Results as presented in Figure 2.14 (left) can be used to derive the maximum access queue size that can be implemented such that the expected additional call set-up time (access time) is less than an operator-specified service requirement of, say, $\alpha_{\max}$ seconds. Although in practice a 90% percentile of the access time would be a more appropriate measure to determine the optimal access queue size, the expected access time given by our model provides a useful and analytically obtainable first-order indication. As illustrated in Figure 2.14 (right) for $\alpha_{\max} = 4$ seconds, the maximal access queue size strongly depends on the data traffic load. Under very low data traffic loads, the expected access time may be sufficiently low even with an infinite access queue size, which is the case in Figure 2.14 (left) for the cells with $14, 21$ or 28 traffic channels, where the access time converges to a maximum below $\alpha_{\max}$. As the data traffic load becomes heavier, the number of positions in the access queue must be reduced in order to meet the $\alpha_{\max}$ requirement, which causes an additional indirect increase in the data call blocking probability, aside from the direct and obvious effect of the heavier load. The observation that for lighter data traffic loads the maximum allowable access queue size tends to be super-linear in the capacity is due to the statistical multiplexing effect described above, while for very heavy data traffic loads the multiplexing gain diminishes and the relation becomes linear, as expected. Note that in Figure 2.14 (right) the curve for $C_{\text{total}} = 14$ does not decrease between data traffic loads of $0.7 \cdot 12.837$ and $0.8 \cdot 12.837$ Erlang due to the discretisation effect: for each Q_a there is a *range* of data traffic load values for which this access queue size is optimal.

2.8.5. CONDITIONAL EXPECTED SOJOURN TIMES

In Corollary 2.2 it was stated that the conditional expected transfer times $\widehat{\boldsymbol{\tau}}(x)$ are asymptotically linear in the data call size x. Since both the expected access time $\mathbf{T}_{a,\text{data}}$ and the transition probability distribution Ψ are independent of the call size, the conditional expected *sojourn* time $\mathbf{T}_{\text{data}}(x)$ must be asymptotically linear in x as well (see (2.2) and (2.3)). For default values of all parameters but ρ_{data}, which is taken from $\{0.1, 0.5, 1.0\} \cdot \rho_{\text{speech}}$, Figure 2.15 below demonstrates the convergence of the exact expected sojourn times to the derived asymptotes. For each value of ρ_{data} in the left chart, the asymptote is the dashed line with the small open markers that almost coincides with the exact curve for $\rho_{\text{data}} \in \{0.1, 1.0\} \cdot \rho_{\text{speech}}$, while it more visibly deviates from the exact curve for $\rho_{\text{data}} = 0.5 \cdot \rho_{\text{speech}}$. The range of x values considered is from 0 to $100\, r_{\text{data}}$ kbits, the latter value corresponding to the 99% percentile of the data call size distribution. Although in this example the exact expected sojourn time values converge to the asymptote from below, this observation does not hold in general, as we learned from other experiments with different parameter settings.

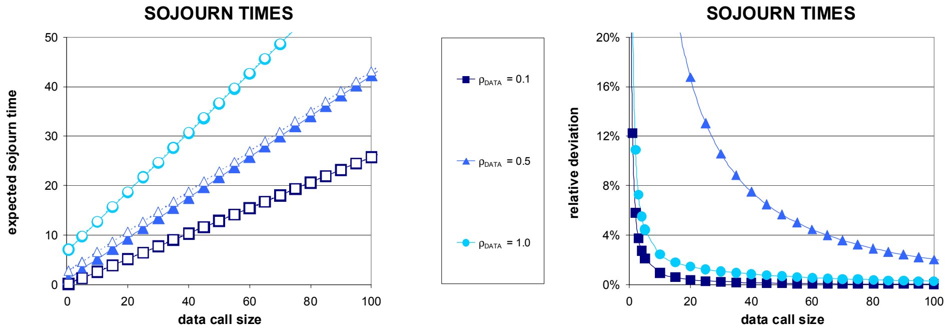


Figure 2.15 Conditional expected sojourn times: convergence of conditional expected sojourn times versus the data call size.

Furthermore, the right chart indicates that the speed of convergence, which is expressed as the relative deviation between the exact and the approximate expected *transfer* time, appears to be lowest for moderate data loads, which we intuitively expect to hold in general. The reason for using the *transfer* time rather than the *sojourn* time to determine the speed of convergence is that we do not want the different values of the expected access time to distort the comparison. In general, the speed of convergence is predominantly determined by the second-largest eigenvalue of the infinitesimal generator $\mathcal{B}^{-1}\mathcal{Q}^{\star}$, as can be seen from the proof of Corollary 2.2. It is extremely difficult to obtain analytical insight into the relation between the eigenvalues of $\mathcal{B}^{-1}\mathcal{Q}^{\star}$ and the model parameters.

An intermediate result in the determination of the conditional expected sojourn times $\mathbf{T}_{\text{data}}(x)$ is given by the $\widehat{\sigma}_{s,d}(x)$, the conditional expected sojourn time of an admitted data call of size x arriving in system state $(s, d) \in \mathbb{S}_{+}$ (recall that d includes the new call). This result may be very useful as a feedback information service to the caller. Figure 2.16 presents an illustrative example of $\widehat{\sigma}_{s,d}(x)$ versus $(s, d) \in \mathbb{S}_{+}$ for an

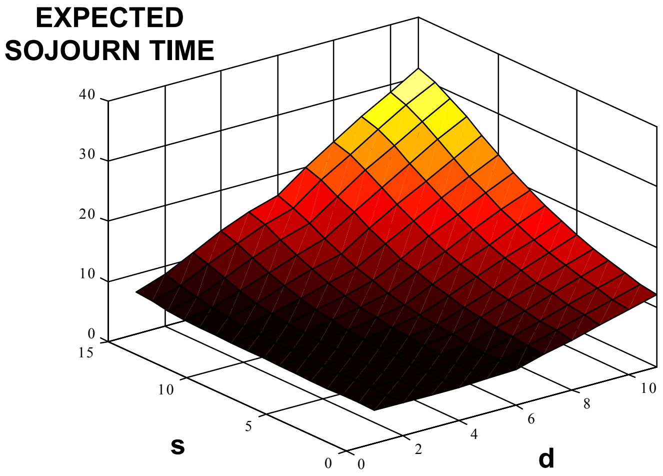


Figure 2.16 Conditional expected sojourn times: conditional expected sojourn time versus system state at arrival.

admitted data call of size 320 kbits, given the default parameter settings. The figure supports the intuition that $\widehat{\sigma}_{s,d}(x)$ is increasing in both s and d, i.e. that the expected conditional sojourn time is longer as the data call finds the cell to be more congested upon its arrival. The numerical example further illustrates that the expected sojourn time of a data call is most sensitive to a change in the number of *data* calls, since an additional active data call claims at least as many traffic channels as a speech call. Note the abrupt change in slope at $d = d_{\max}(s) - Q_a = 6$, indicating an increased sensitivity of the expected sojourn time of a queued data call with respect to the number of data calls queued ahead of it.

2.9. CONCLUDING REMARKS

We have presented an extensive analytical performance evaluation of a class of fair channel sharing policies in an integrated GSM/HSCSD network with a finite access queue for data call requests that cannot be served immediately upon arrival. Markov chain analysis has been applied to obtain simple performance measures such as channel utilisation, speech and data call blocking probabilities and the average data call access, transfer and sojourn times. Furthermore, using differential equations, an analytical expression has been derived for the expected sojourn time of a data call, conditional on its size, indicating that the sojourn time is asymptotically proportional to the call size and hence the proposed policies provide fairness with respect to various data call sizes. As a valuable intermediate result in the analysis, the conditional expected sojourn time of an admitted data call is obtained, given the system state at arrival, which may serve as an appreciated feedback information service to the data source.

Four typical channel sharing policies within the given class have been specified for numerical evaluation. Among these the SHARE-RESERVE policy has been argued to be most promising. Under this policy, an operator-specified number of traffic channels is reserved for data transfer only, while the remaining channels are shared by speech and data calls. On the shared channels speech calls are strictly prioritised, with service preemption over data calls. At any time, the active data calls fairly share all available channels up to the terminals' multislot capabilities, with channel assignments that are dynamically adapted to either utilise freed capacity or to support newly admitted data calls.

Expecting that a mobile network operator is likely to be rather hesitant to degrade its speech service when entering the data market, the SHARE-RESERVE policy performs best when the data traffic load grows from light to moderate or heavy, given its generally high channel utilisation and the protection it offers to the speech users, independent of an increase in the data traffic load. The desired trade-off between the GRade and QUality OF SErvice measures can be achieved by adapting the reservation level to the data traffic load. Initially, only for very light data traffic loads, it seems a waste to reserve any channels for data transfers, so an operator is better off with the SHARE policy, sharing all channels and reducing the preferential treatment of speech calls such that data calls can only be downgraded to $\beta^{\min}$ traffic channels, e.g. $\beta^{\min} = 1$. Since it is most robust against a data traffic load increase, the SHARE-RESERVE policy has been selected for a further numerical investigation, presented to obtain insight in the performance effects of the various system and policy parameters, and to illustrate the sojourn time expectations that may be fed back to a data caller.

CHAPTER 3

SERVICE INTEGRATION IN A GSM/GPRS NETWORK

THE development and deployment of the General Packet Radio Service is instigated by the initially separate yet now converging growth both in mobile speech telephony and data communications. Moreover, mobile service providers are targeting the consumer market of multimedia messaging, i-mode, downstreaming audio and video clips and foresee a demand for video telephony. Aiming for higher data rates, flexible channel assignments and enhanced resource efficiency, the introduction of GPRS into existing GSM networks is a key step in the current evolution in mobile networks, which is characterised by a transition from circuit-switched speech-oriented networks to integrated circuit- and packet-switched multi-service networks.

As technological innovations alone are undeniably insufficient to achieve the QOS and efficiency objectives, the current chapter develops and analyses a generic performance model for the evaluation and optimised deployment of capacity allocation mechanisms in an integrated services GSM/GPRS network. In an attempt to capture the relevant aspects of the principally expected services, the distinctly characterised (circuit-switched) speech and (packet-switched) video, high- and low-priority data services are incorporated in the analysis. The considered model and analysis extend those developed in Chapter 2 e.g. in the inclusion of a new service type and the capability to deal with the absence of a non-zero channel assignment guarantee, which is intrinsic to packet-switched networks.

The outline of the chapter is as follows. The related literature is discussed in Section 3.1. Subsequently, Section 3.2 states the principal contributions of this chapter. Section 3.3 then gives an overview of the GSM/GPRS system and describes the relevant aspects for our investigations. The mathematical framework is defined in Section 3.4 in the form of two separate models followed by a discussion on the generality of the framework, which allows diverse model considerations. Subsequently, Section 3.5 presents a basic performance analysis that is based on direct application of the equilibrium distribution of the constructed Markov chain. A conditional performance analysis is given in Section 3.6, deriving expressions for the video and data QOS, conditional on their service requirement and the system state upon arrival. An extensive numerical study is presented in Section 3.7 in order to provide quantitative insight in the relevant GOS and QOS trade-offs, while Section 3.8 ends this chapter with some concluding remarks.

3.1. LITERATURE

The literature review below concentrates on two aspects of the presented investigation. We first discuss some references that focus on the performance evaluation of integrated services (GSM/)GPRS networks, applying either simulation or analytical techniques, followed by a review of some studies that assess the feasibility of offering real-time services over GPRS networks. Refer to Section 2.1 for an extensive review of the stochastic analysis of integrated services models in a generic or wireline setting.

3.1.1. SERVICE INTEGRATION IN GSM/GPRS NETWORKS

The initial performance evaluation studies of (GSM/)GPRS networks that can be found in the literature typically focus on (generally single cell) data-only networks, using system-level simulations in order to *(i)* investigate the impact of the system and environment parameters such as data traffic load, call size distribution and channel availability on performance measures such as call blocking, delay and throughput (see e.g. [37, 39, 41]); or *(ii)* compare different scheduling algorithms (e.g. First-In First-Out, Round Robin, Earliest Deadline First, Static Priority Scheduling) to handle data traffic (see e.g. [4, 120, 180, 205]). A multiple cell network is considered by [120] in order to model frame errors due to interference. Pure or slightly modified versions of the Round Robin and Earliest Deadline First disciplines are generally considered most promising, although the latter is also expected to be rather complex from an implementation viewpoint [180]. Integrated speech (GSM)/data (GPRS) networks with a dynamic radio resource allocation algorithm have been studied via simulation by e.g. [27, 50, 126], where the Grade- (GOS: speech) and Quality of Service (QOS: data) are determined as a function of the speech and data traffic loads, as well as the number of dedicated data traffic channels. Based on single cell GSM/GPRS simulations featuring both up- and downlink transfer of multiple data traffic types, [134] determines the combinations of speech and data traffic loads that are feasible under given GOS and

QOS requirements. A comparison of GPRS's four channel coding schemes is reported in [104], presenting the relation between throughput and radio link quality.

Analytical performance studies in this field are rare. Using a simpler version of the model and analysis presented in this chapter, the performance of a fully segregated and a hybrid radio resource sharing scheme has been evaluated and analytically compared in [151], in order to quantify the capacity gain that can be achieved when utilising the idle periods between speech calls by filling these gaps with packet data. In [14] a single cell in a GSM-based network is studied serving speech, scalable video and elastic data calls, using straightforward Markov chain analysis to obtain channel utilisation and call blocking probabilities. Finally, [82] models the evolution of a GSM/GPRS cell as a single server queue in a Markovian environment, assuming a two-state Markov modulated Poisson process to describe the packet data arrival process. Steady-state performance measures are derived from the constructed Markov chain's equilibrium distribution, which is obtained using matrix-geometric techniques (see e.g. [172]).

3.1.2. FEASIBILITY OF REAL-TIME SERVICES IN GPRS NETWORKS

With respect to the feasibility of real-time services in GPRS networks that are more bandwidth-demanding than speech telephony, we refer to studies reported in [47, 80, 206]. [47] investigates the possibility of embedding audio streaming in an E-GPRS network at 16 or 32 kbits/s, proposing a set of performance enhancement techniques to overcome the problems of high error rates and bursty error patterns. [80] examines real-time ISO (International Organisation for Standardisation) MPEG-4 (Motion Pictures Expert Group) video services over GPRS, noting that unlike ITU's (International Telecommunication Union) H.263 video codec, the MPEG video codec was specifically designed with mobile networks in mind, featuring an advanced suite of error resilience tools. The reported experiments indicate that for five frames per second, a minimum acceptable video quality can be achieved using two CS-2 or CS-3 (see Section 3.3.2 below) traffic channels at SIR values above 11 dB and 18 dB, respectively. In [206] both voice-over-IP and H.263 coded video have been simulated over GPRS as a bearer service. Simulation experiments assuming assignments of two CS-2 traffic channels per video call demonstrated that the hierarchical (scalable) coding mode of the H.263 video codec, which allows to discard frames according to their relevance for the overall picture quality, produces a better image quality in case of (radio) network congestion, compared to traditional single layer coding.

3.2. CONTRIBUTION

The principal contribution of this chapter is the development and analysis of a *generic model* for performance evaluation, parameter optimisation and dimensioning in a GSM/GPRS network. The model enables *analytical evaluation* for an integration scenario with speech and elastic video and data services, potentially offered in distinct priority classes. The video and data QOS is expressed in the experienced (conditional) expected video throughput and sojourn times, respectively. The considered set of services cover the principal characteristics specifying the different traffic/QOS classes that are standardised for integrated services networks (see e.g. [105]). Although a wider variety of models can be designed and analysed within the generic framework, the analytical and numerical results are presented for the SVD model integrating speech, video and data calls, and for the SHL model, integrating speech, high- and low-priority data calls (e.g. delay-sensitive WWW browsing sessions and delay-tolerant e-mail transfers). An extensive *numerical evaluation* is included to demonstrate the merits of the studied generic model and performance analysis, and to provide insights in the performance trade-offs involved in balancing the various system parameters.

3.3. GENERAL PACKET RADIO SERVICE

The General Packet Radio Service [70, 72, 73, 74] is a PHASE 2+ upgrade of the widely deployed GSM system, enhancing resource efficiency and offering higher data rates primarily due to its packet-switched character. In principle, the GPRS specifications allow data rates of up to $8 \times 21.4 = 171.2$ kbits/s, although the actual assignments are subject to the data traffic load, the terminals' multislot capabilities and the required degree of forward error coding, which is related to the experienced radio link quality. GPRS is suitable to handle a variety of services with different (and varying) bandwidth requirements, including e-mail, WWW browsing, file transfer, audio/video streaming and video telephony [25, 37, 39, 92, 122].

3.3.1. NETWORK ARCHITECTURE

Figure 3.1 presents the network architecture of an integrated GSM/GPRS network. GSM speech calls are handled in a circuit-switched manner, i.e. an end-to-end dedicated connection is established between the Mobile Station and another mobile or fixed telephone, via the serving Base Transceiver Station, a Base Station Controller, optionally a Mobile Switching Center, and finally a Gateway Mobile Switching Center, which functions as the interface between the mobile network and the fixed network (PSTN (Public Switched Telephone Network) or ISDN (Integrated Services Digital Network)). In order to integrate GPRS into the existing GSM architecture, two new network entity types must be installed: a Gateway GPRS Support Node (GGSN), which is a fairly simple router that forms the interface with external Packet Data (e.g. IP or X.25) Networks, and a Serving GPRS Support Node (SGSN), which is responsible for the delivery of packets to and from the MSs within its service area. The SGSN tasks include packet routing and transfer, mobility management, authentication and charging functions. Each BSC is to be upgraded with a Packet Control Unit, which controls the packets' transmission over the radio interface, via the base transceiver station.

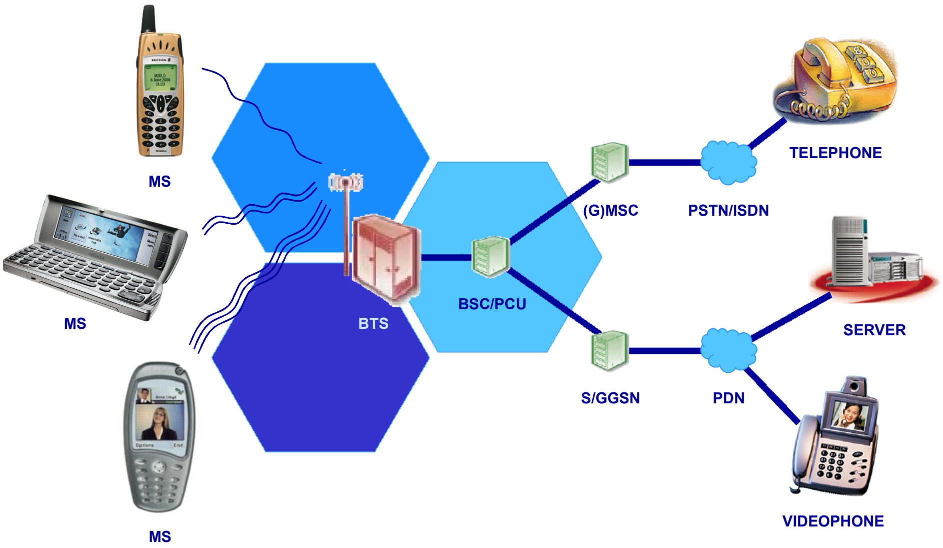


Figure 3.1 GSM/GPRS network architecture: the illustration shows example speech, video and data calls maintained over circuit- and packet-switched connections, respectively.

Consider the case of a mobile-terminated data call. The remote host sends its data packets via an external packet data network, e.g. the Internet, to the GGSN, which subsequently routes the packets over the intra-PLMN (Public Land-Mobile Network) IP backbone to the SGSN. This router is connected to the target PCU. Packet buffering takes place in both the SGSN and the PCU. Since in the end-to-end path the wireless segment is typically the bottleneck, and given the anticipated traffic asymmetry, we focus on resource sharing in the *downlink* of the *radio interface*.

3.3.2. RADIO INTERFACE

Aside from the logical *control* channels that may be shared by GSM and GPRS calls, the remaining pool of *traffic* channels can be either dedicated to or dynamically shared by circuit-switched GSM calls and packet-switched GPRS calls. The basic transmission unit on a GPRS traffic channel is called a radio block which consist of four time slots in four consecutive TDMA frames. As every 13th TDMA frame is left idle, the mean transmission time per radio block is 20 ms, which is also the time scale for data scheduling. Each radio block contains 456 bits, due to GSM's Gaussian Minimum Shift Keying (GMSK) modulation scheme, while the actual payload depends on the degree of forward error correction coding that is applied. The GPRS standards specify four convolutional coding schemes with different code rates, protection levels and corresponding throughputs (see Table 3.1). Only CS-1 and CS-2 will be available initially due to bit rate limitations on the BTS-BSC interface.

Table 3.1 GPRS coding schemes: four distinct coding schemes have been specified with different error correction capabilities and different effective throughputs.

	CS-1	CS-2	CS-3	CS-4	
code rate	1/2	$\approx 2/3$	$\approx 3/4$	1	
throughput	9.05	13.4	15.6	21.4	kbits/s

Equivalent to the limitations for the HSCSD terminals discussed in Section 2.3, the (downlink) multislot capability $\beta^{\max}_{\text{GPRS}}$ is terminal-dependent and is typically 2, 3 or 4. As noted earlier for the HSCSD standard, traffic channels assigned to a given GPRS terminal must be adjacent on a single frequency, although this practical requirement is neglected in our analysis (see also Section 2.4.1). Another limitation exists regarding the maximum number of data calls that can be assigned to any given downlink traffic channel. For technical reasons, this number is limited to nine in a typical GPRS implementation, while in our radio resource sharing schemes the operator is able to set this maximum at a lower value.

3.3.3. QUALITY OF SERVICE

In view of the diversity of the expected services, support of different QOS classes is an important feature of GPRS. The call-specific QOS profile consists of four distinct attributes. The *service precedence* denotes the relative priority of a service and plays a role when e.g. unanticipated congestion requires some calls to be terminated prematurely. The *reliability* indicates the transmission characteristics required by an application in terms of e.g. the maximum allowed packet loss probability. The *delay* attribute defines a maximum for both the average delay and the 95th delay percentile, which refer to the delays experienced within the GPRS network only. Finally, the *throughput* attribute specifies the peak and mean bit rates.

Based on these attributes, the QOS profile of a call is negotiated for each session during a phase known as PDP (Packet Data Protocol) context activation, based on the desired QOS and the current availability of resources. Aside from transferred data volume and the service type, also the chosen QOS profile may be incorporated in the charged price.

While the GPRS specifications define the above QOS metrics, it is up to the system manufacturer to design and implement the necessary mechanisms that enable QOS provisioning in order to meet the negotiated QOS and prevent that congestion degrades performance intolerably. The network operators in turn have the freedom of setting the appropriate parameters associated with the available QOS control mechanisms. In an integrated services setting, the presented study focuses on three such mechanisms: Radio Resource Reservation (partitioning the available pool of resources), Call Admission Control and packet scheduling.

3.4. MODELS

This section defines the framework for the performance analysis of a GSM/GPRS network that integrates *speech*, *video*, and *data* services, differentiating between a *high*- and a *low-priority* data class. These three services have been selected for the fundamental differences in the service characteristics, in the sense that speech calls request a fixed channel assignment for a given duration, (scalable) video calls can handle a varying channel assignment, also for a given duration, while data calls can also handle a varying channel assignment which however directly influences the call duration. In this light, we note that the specified video call characteristics in their generic form also capture e.g. downlink streaming of stored audio or video information (clips) or service, in addition to two-way conversational video communications.

Although the presented model is in principle suitable to analyse a network that integrates all four service types, potentially with an even broader distinction in priority classes, we choose to split this more generic model into two distinct models, to enhance the transparency of the performance analysis: the SVD model which integrates *speech*, *video* and *data* services and the SHL model which integrates *speech*, *high-* and *low-priority data* services (with QOS differentiation).

3.4.1. TRAFFIC MODEL

This subsection describes the common aspects of the SVD and SHL models regarding the *service characteristics* of the speech, video and data calls. The characteristics of the speech and data services are similar or even identical to those in Chapter 2. For reasons of analytical tractability, the speech and video call durations as well as the data call sizes are assumed to be exponentially distributed. A sensitivity analysis of the data QOS regarding the impact of the data call size variability will be presented in Chapter 4. All arrival processes and call duration (speech, video) or size (data) distributions are assumed to be mutually independent.

SPEECH SERVICE: Speech calls arrive according to a Poisson process with arrival intensity λ_{speech} and have an exponentially distributed duration with mean $1/\mu_{\text{speech}}$. A speech call requires a fixed assignment of one traffic channel, and the speech traffic load is given by $\rho_{\text{speech}} \equiv \lambda_{\text{speech}}/\mu_{\text{speech}}$.

VIDEO SERVICE: Video calls arrive according to a Poisson process with arrival intensity λ_{video} and have an exponentially distributed duration with mean $1/\mu_{\text{video}}$. Video calls are modelled as continuous real-time streams that are *scalable* in the sense that the number of assigned traffic channels and thus the *audio/image quality* is adaptive to the varying network load. Scaling is assumed to adhere to any channel reassignment (up- or downgrade) ideally and instantaneously. As these up- and downgrade events occur on a time scale of typically seconds, i.e. the time scale at which call arrivals and departures occur, the assumption of instantaneous adjustment of the source transfer rate is expected to only weakly influence the numerical results.

In order to guarantee a certain video quality, each video call must be assigned at least a minimum of $\beta^{\min}_{\text{video}} \in [0, \beta^{\max}_{\text{GPRS}}]$ traffic channels, corresponding to a bit rate of $r_{\text{video}}\, \beta^{\min}_{\text{video}}$ kbits/s, with r_{video} the fixed information bit rate (in kbits/s) per video traffic channel (e.g. $r_{\text{video}} = 13.4$ kbits/s under GPRS channel coding scheme CS-2). Although effectively the video traffic load is obviously influenced by $\beta^{\min}_{\text{video}}$, the definition of choice is $\rho_{\text{video}} \equiv \lambda_{\text{video}}/\mu_{\text{video}}$, as it allows the absence of an absolute QOS guarantee ($\beta^{\min}_{\text{video}} = 0$).

DATA SERVICE: Data calls arrive according to a Poisson process with arrival intensity λ_{data}. In the SHL model with QOS differentiation for the data service, a fraction ξ of data calls is of high-priority while the complementary fraction $1 - \xi$ is of low-priority. As in the previous chapter, a data call is assumed to be the downlink transfer of a file with an exponentially distributed length. Given a fixed information bit rate of r_{data} kbits/s per data traffic channel, the nominal data call transfer time is also exponentially distributed with mean denoted $1/\mu_{\text{data}}$, given the exclusive use of a single traffic channel. The data call length is expressed in units of r_{data} kbits and thus has mean value $1/\mu_{\text{data}}$, which corresponds to $r_{\text{data}}/\mu_{\text{data}}$ kbits. The data traffic load is given by $\rho_{\text{data}} \equiv \lambda_{\text{data}}/\mu_{\text{data}}$ (expressed in Erlang), which in the SHL model consists of $\rho_{\text{high}} \equiv \xi\, \rho_{\text{data}}$ Erlang high-priority data traffic and $\rho_{\text{low}} \equiv (1 - \xi)\, \rho_{\text{data}}$ Erlang low-priority data traffic.

Data calls are assumed to be *elastic* in the sense that they are delay-tolerant and can handle varying channel assignments. The number of traffic channels that can be assigned to a data call is limited to the technical maximum $\beta^{\max}_{\text{GPRS}}$ (see Section 3.3). Unlike for video calls, no minimum transfer rate is guaranteed, hence the admitted data calls' assignment may potentially be downgraded to zero. In practice this implies that the downloaded data is temporarily buffered for later transfer. We further assume that for a given call there is at any time sufficient data available in the PCU/SGSN buffers to be carried on the dynamically assigned channels.

Observe from the service specifications above that the key difference between video and data calls is the impact of the channel assignment on the perceived QOS and on the calls' presence in the system: for the video service, the channel assignment influences the perceived audio/image quality while it does not affect the autonomously sampled video call duration, while for the data service, the channel assignment affects the rate at which the file is transferred and thus the data call's sojourn time (see also [190]).

With regard to the final assumption made for the data service, we note that in a real network end-to-end flow control is executed at the TCP layer, which controls the rate at which the PCU/SGSN buffers are fed, in order to limit the amount of data in transit. Assuming that the wireless segment is the primary bottleneck in the end-to-end connection, we argue that the variability at which TCP feeds the buffer is induced by and hence in direct correspondence with the variability of the data transfer rate over the air interface, which supports the above assumption. In fact, the implemented Processor Sharing channel sharing scheme is not only attractive due to its inherent fairness property, but it also serves as a convenient modelling abstraction for an inclusion of the (idealised) effects of TCP flow control in the sense of (instantaneous) adjustment of the transfer rate in accordance with the traffic congestion level. See the intermezzo below for a further elaboration on the impact of TCP on data performance in wireless networks.

INTERMEZZO: ON THE IMPACT OF TCP IN A WIRELESS NETWORK

Unless the data call is processed over an end-to-end circuit-switched connection, end-to-end error and congestion control are typically executed at the TCP transport layer operating on top of the IP network layer. In several publications (e.g. [186, 193]), potential problems have been identified at the transport layer that arise in the case of a wireless access network. These are a consequence of the design premise that TCP was intended for relatively fast and reliable fixed networks, rather than for slow and unstable radio networks. Regarding *error control*, the relatively high frame error rates and corresponding link level delays that are common in the radio interface can easily induce TCP to generate unnecessary retransmissions. TCP may even confuse excessive delays with a connection loss, forcing the application to make an expensive reconnection. Furthermore, TCP's *congestion control* may falsely interpret such link level delays as a symptom of congestion, whereas it may very well be caused by link level retransmissions of erroneous frames due to fading. In response, transmission windows and retransmission time-outs may be adjusted, leading to a potentially inappropriate flow reduction. As a consequence, data throughputs and resource efficiency are reduced, a highly undesirable effect in the radio interface, given its intrinsically scarce capacity.

Although little practical experience exists regarding these matters, due to the fact that mobile data communications is still relatively immature, a number of theoretical performance studies and proposals for TCP improvement can be found in the literature.

Aside from physical layer solutions of increased forward error correcting coding to lower the frame error rate at the cost of a reduced throughput, higher-layer solutions for the anticipated problems attempt to fool TCP by hiding the lossiness of the wireless link (e.g. [10, 186, 193]). In contrast to the concerns that triggered these studies, the performance analysis of standard TCP for GSM/GPRS networks presented in [169] demonstrates that TCP and GPRS's ARQ (Automatic Repeat reQuest) mechanism are well harmonised, as the ARQ scheme is appropriately designed to ensure that TCP observes just packet delays rather than packet losses.

The model considered in this chapter ignores frame errors on the data traffic channels and hence any experienced (queueing) delay is indeed caused by radio interface *congestion* only (not due to link level retransmissions). Such delays therefore suitably induce TCP to slow down the source rate. As argued above, under an ideal TCP feedback mechanism the variability at which TCP feeds the PCU/SGSN data buffers, is synchronised with the variability of the data transfer rate over the air interface, which supports the assumption that the buffers are never empty as long as the file is not fully transferred (ignoring TCP slow start effects). The PS service discipline applied in our model is in fact regularly selected to specifically model the TCP-induced effects of fair resource sharing (e.g. [16, 31, 128, 164, 175]).

3.4.2. SYSTEM MODEL: COMMON ASPECTS

Although the call handling procedures are substantially different for the SVD and SHL models, we introduce here the common system model aspects, as well as the applied notation. Precise descriptions of the model-specific call handling procedures are given in the subsequent subsections.

Each model focuses on a single cell in a GSM/GPRS network, serving circuit-switched speech on GSM bearers and packet-switched video and/or data calls on GPRS bearers. Denote with C_{total} the number of traffic channels available in the cell. In order to govern the sharing of this cell capacity by the different service types, the model-specific *channel sharing* schemes split the pool of C_{total} traffic channels into distinct subpools with different purposes. Call handling is further defined by call

Admission Control schemes, specified by the functions

$$\begin{aligned} &\text{SVD:} \quad s_{\max}(s,v,d), \quad v_{\max}(s,v,d), \quad d_{\max}(s,v,d), \\ &\text{SHL:} \quad s_{\max}(s,h,\ell), \quad h_{\max}(s,h,\ell), \quad \ell_{\max}(s,h,\ell), \end{aligned}$$

for the SVD and SHL model, whose states are denoted (s, v, d) and (s, h, ℓ), with s, v, d, h, and ℓ the number of speech, video, data, high- and low-priority data calls present in the system, respectively. An incoming call of a certain type is blocked if and only if upon arrival the present number of calls of that type equals the corresponding (state-dependent) maximum. All blocked calls are cleared from the system.

An admitted call is assigned a number of channels defined by the *channel assignment* (or scheduling) schemes

$$\begin{aligned} &\text{SVD:} \quad \beta_{\text{speech}}(s,v,d), \quad \beta_{\text{video}}(s,v,d), \quad \beta_{\text{data}}(s,v,d), \\ &\text{SHL:} \quad \beta_{\text{speech}}(s,h,\ell), \quad \beta_{\text{high}}(s,h,\ell), \quad \beta_{\text{low}}(s,h,\ell), \end{aligned}$$

which prescribe for each model and system state the *expected* number of channels that are assigned to each call type. The reason we write 'expected' is due to the fact that in practice at any time a given channel can be assigned to a single call only. Hence if, for instance, $\beta_{\text{data}}(s, v, d)$ is a fractional number, e.g. 2.7, this implies that in state (s, v, d) some (randomly selected) data calls are assigned 2 channels while 3 channels are assigned to the other data calls in transfer. As in a GPRS network the assignment of channel capacity over the present calls is done with a relatively small heartbeat of 20 ms, we may model the implementable Round Robin discipline by a Processor Sharing discipline, its idealised continuous-time, analytically more tractable equivalent. The channel assignment schemes incorporate the limitations imposed by the GPRS terminals' multislot capability.

3.4.3. SPEECH/VIDEO/DATA MODEL

The SVD model concentrates on a single cell in a GSM/GPRS network, serving circuit-switched speech calls on GSM bearers, and packet-switched video and data calls on GPRS bearers (see Figure 3.2).

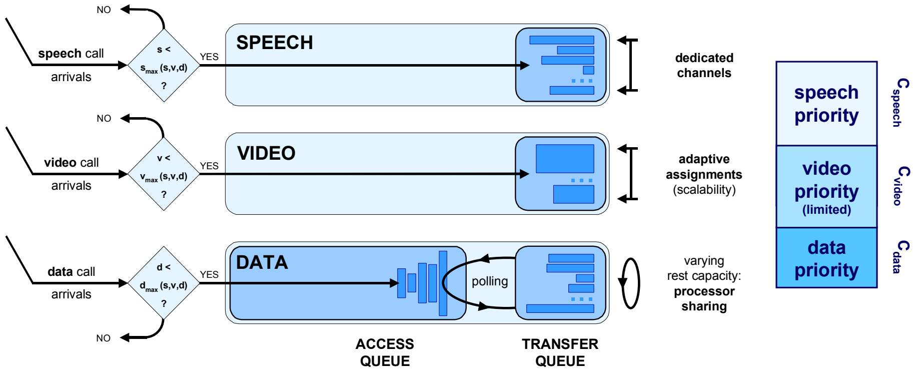


Figure 3.2 The SVD model integrates speech, video and data calls according to specific call handling procedures.

In the SVD model, the proposed *channel sharing* scheme splits the pool of C_{total} traffic channels into three distinct subsets. Although a variety of alternatives can be defined and analysed within the same generic analytical model, the considered scheme is merely a suggestion for channel sharing that we deem sensible, since it establishes some form of capacity reservation for all service types, while still providing a high channel utilisation though varying elastic call assignments (see also the comparison in Section 2.8.1). We note that the design of a channel sharing scheme is typically done by the radio network vendor, while the network operator sets the parameters in accordance with its traffic expectations and policy regarding (differentiated) GOS and QOS.

Among the C_{total} traffic channels, C_{speech} channels are reserved for speech calls with preemptive priority, i.e. video and data calls may use these channels whenever they are unused by the speech service, but must free them immediately once needed to support newly admitted speech calls. Within this set of C_{speech} channels, video calls are treated with strict preference over data calls. C_{video} channels are shared by all call types with preference for speech and video calls. Video calls must downgrade their assignment (potentially down to $\beta_{\text{video}}^{\text{min}}$ channels) only in support of newly admitted speech or video calls. It is in this channel pool that video calls are protected, and must thus find their minimum assignment of $\beta_{\text{video}}^{\text{min}}$ channels, as channels grabbed elsewhere may have to be released again in favour of newly admitted calls. Data calls may utilise the capacity that cannot be assigned to the preferred speech or video calls. Lastly, $C_{\text{data}} \equiv C_{\text{total}} - C_{\text{speech}} - C_{\text{video}}$ channels are reserved for data calls with preemptive priority, i.e. video calls can grab additional channels from this pool due to their scalability property but only if such channels would otherwise be idle, i.e. if $d_t(s,v,d)\,\beta^{\max}_{\text{GPRS}} < C_{\text{data}}$, with $d_t(s,v,d)$ the number of active data calls (see below). Speech calls are prohibited to use these channels. Although in principle each of these parameters may be set to zero, support for video services requires that $C_{\text{video}} \geq \beta^{\min}_{\text{video}}$.

As we will see, the system evolution can be modelled as a continuous-time Markov chain $(S(t), V(t), D(t))_{t\geq 0}$ where $S(t)$, $V(t)$ and $D(t)$ are defined as the number of speech, video and data calls, respectively, that are present at time t. The system states are denoted (s,v,d) with state space $\mathbb{S}$.

The Call Admission Control and channel assignment schemes of the SVD model are defined as follows. A *speech* call is blocked if and only if upon arrival no traffic channel is, or can be made available to support the call, i.e.

$$s_{\max}(s,v,d) \equiv \left\lfloor C_{\text{speech}} + C_{\text{video}} - \beta^{\min}_{\text{video}} v \right\rfloor,$$

while an admitted speech call is assigned a single traffic channel for its entire duration:

$$\beta_{\text{speech}}(s,v,d) \equiv 1.$$

A *video* call is blocked if and only if upon arrival the minimum assignment of $\beta^{\min}_{\text{video}}$ channels cannot be made available to support the call, i.e.

$$v_{\max}(s,v,d) \equiv \left\lfloor \frac{C_{\text{video}} - \max\{s - C_{\text{speech}}, 0\}}{\beta^{\min}_{\text{video}}} \right\rfloor,$$

with $\max\{s - C_{\text{speech}}, 0\}$ the number of shared channels that are in use by speech calls. In order to determine the expected channel assignment $\beta_{\text{video}}(s,v,d)$ of an admitted video call in system state (s,v,d), first note that the number of channels available for video transfer is given by $\max\{C_{\text{data}} - \beta^{\max}_{\text{GPRS}}\, d_t(s,v,d), 0\} + (C_{\text{video}} + C_{\text{speech}} - s)$, with $d_t(s,v,d)$ the number of active data calls (see below). The first part of this expression indicates the number of channels in the C_{data} pool that is available while the second part gives the number of available channels in the joint $C_{\text{video}} + C_{\text{speech}}$ pool. The available channels are then distributed as evenly as possible over the present video calls:

$$\begin{aligned}&\beta_{\text{video}}(s,v,d)\\ &\quad\equiv\quad \min\left\{\beta_{\text{GPRS}}^{\max},\frac{\max\left\{C_{\text{data}}-\beta_{\text{GPRS}}^{\max}\,d_t(s,v,d),0\right\}+\left(C_{\text{video}}+C_{\text{speech}}-s\right)}{v}\right\},\end{aligned}$$

respecting the terminals' multislot capabilities. It is readily verified that the Call Admission Control schemes of speech and video calls guarantee that $\beta_{\text{video}}(s,v,d)\geq\beta_{\text{video}}^{\min}$ for all possible states $(s,v,d)\in\mathbb{S}$.

Data calls are logically organised in two distinct queues. The *transfer queue* can hold up to Q_t data calls. In system state (s,v,d), denote with $d_t(s,v,d)\equiv\min\{d,Q_t\}$ the number of data calls in the transfer queue. A data call in transfer is also referred to as an *active* data call, even if at times of congestion the network may exploit its delay tolerance and temporarily downgrade its channel assignment to zero, depending on the call handling parameters. As in the model of Chapter 2, a FIFO *access queue* is maintained to hold admitted data calls that cannot enter the transfer queue immediately. Denote with Q_a the number of data call requests the access queue can store. In system state (s,v,d) denote with $d_a(s,v,d)\equiv d-d_t(s,v,d)=\max\{0,d-Q_t\}$ the number of data calls in the access queue. Upon termination of an active data call, the data call at the head of the access queue is immediately polled into the transfer queue, where it immediately shares in the assignment of available traffic channels. Hence

$$d_{\max}(s,v,d)\equiv Q_a+Q_t,$$

where we write $d_{\max}(s,v,d)$ for uniformity of presentation and to indicate that one may define a *dynamic* Call Admission Control threshold (as e.g. for the HSCSD data calls in Chapter 2), although in the presented model it is independent of the system state. At any time, the channel capacity that is available for the data service is fairly shared by all active data calls according to a PS service discipline, i.e.

$$\beta_{\text{data}}(s,v,d)\equiv\min\left\{\beta_{\text{GPRS}}^{\max},\frac{C_{\text{total}}-s-\beta_{\text{video}}(s,v,d)\,v}{d_t(s,v,d)}\right\},$$

respecting the terminals' multislot capabilities. By convention, $\beta_{\text{data}}\,(s,v,d)=0$ if there are no active data calls (note that there are no active data calls only if $d=0$).

The data service model is basically a combination of a finite FIFO access queue and a PS transfer queue served at a varying rate due to the speech and video call arrival and termination process. The appropriate setting of system parameters allows an investigation of two extremes as special cases of the presented model: $Q_t = 1$ defines a simple FIFO queue, while $Q_a = 0$ gives a simple PS queue, both under varying service capacity.

The system state descriptors $d_t(s,v,d)$ and $d_a(s,v,d)$, the Call Admission Control thresholds $s_{\max}(s,v,d)$, $v_{\max}(s,v,d)$ and $d_{\max}(s,v,d)$, and the channel assignment functions $\beta_{\text{speech}}(s,v,d)$, $\beta_{\text{video}}(s,v,d)$ and $\beta_{\text{data}}(s,v,d)$, fully specify the studied channel sharing scheme and are thus the only ingredients required for our performance analysis. Note that five parameters are to be set by the network operator, namely C_{speech}, C_{video}, C_{data}, Q_t, and Q_a. All call handling procedures are then explicitly defined.

3.4.4. SPEECH/HIGH-/LOW-PRIORITY DATA MODEL

The second call handling model we define is the SHL model (see Figure 3.3), considering a single cell in a GSM/GPRS network, serving circuit-switched speech calls and packet-switched data calls on GSM and GPRS bearers, respectively. QOS differentiation is with respect to a high- and a low-priority data class.

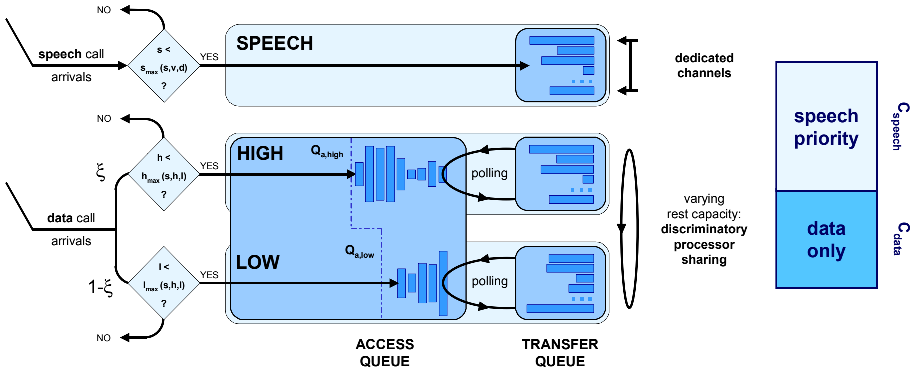


Figure 3.3 The SHL model integrates speech and two classes of data calls according to specific call handling procedures.

In the SHL model, a typically selected and implemented *channel sharing* scheme splits the pool of C_{total} traffic channels into two distinct subsets (see Chapter 2, or also e.g. [50, 82, 126]). C_{speech} channels are reserved for speech calls with preemptive priority, i.e. data calls of both priority classes may use these channels whenever they are unused by the speech service, but must free them immediately once needed to support newly admitted speech calls. The remaining $C_{\text{data}} \equiv C_{\text{total}} - C_{\text{speech}}$ channels are strictly reserved for data calls of both priority classes. With regard to limitations on the parameter settings, we note that whereas $C_{\text{speech}} = 0$ seems senseless for non-zero speech traffic loads as it yields a 100% speech call blocking probability, in contrast, $C_{\text{data}} = 0$ is a perfectly plausible option, especially in cases of low data traffic loads.

As we will see, the system evolution can be modelled as a continuous-time Markov chain $(S(t), H(t), L(t))_{t \geq 0}$ where $S(t)$, $H(t)$ and $L(t)$ are defined as the number of speech, high- and low-priority data calls, respectively, that are present at time t. The system states are denoted (s, h, ℓ) with state space denoted $\mathbb{S}$.

The Call Admission Control and channel assignment schemes are defined as follows. As for the SVD model, a *speech* call is blocked if and only if upon arrival no traffic channel is, or can be made available to support the call, i.e.

$$s_{\max}(s, h, \ell) \equiv C_{\text{speech}},$$

while a fixed single-channel assignment applies to admitted speech calls:

$$\beta_{\text{speech}}(s, h, \ell) \equiv 1.$$

Data calls of each class are logically organised in two distinct types of queues. For each priority class a separate *transfer queue* is maintained that can hold up to Q_t data calls. In system state (s, h, ℓ), denote with $h_t(s, h, \ell) \equiv \min\{h, Q_t\}$ and $\ell_t(s, h, \ell) \equiv \min\{\ell, Q_t\}$ the number of data calls in the high- and low-priority transfer queue, respectively. A data call in transfer is also referred to as an *active* data call. Separate FIFO *access queues* are maintained for each priority class to hold admitted data calls that cannot enter the transfer queue immediately. Denote with Q_a the total number of data call requests the access queues can jointly store. This storage space is to be shared by both priority classes, with $Q_{a,\text{low}} \leq Q_a$ and $Q_{a,\text{high}} \leq Q_a$ the maximum number of low- and high-priority data call requests that can be stored, respectively.

In system state (s, h, ℓ), denote with $h_a(s, h, \ell) \equiv h - h_t(s, h, \ell) = \max\{0, h - Q_t\}$ and $\ell_a(s, h, \ell) \equiv \ell - \ell_t(s, h, \ell) = \max\{0, \ell - Q_t\}$ the number of data calls in the high- and low-priority access queue, respectively. Note that $Q_{a,\text{low}} + Q_{a,\text{high}} > Q_a$ indicates that high- and low-priority data calls compete over space in the access queue. Upon termination of a data call of any priority class, the data call at the head of the corresponding access queue is immediately polled into the proper transfer queue, where it immediately shares in the assignment of available traffic channels. Hence the Call Admission Control thresholds are given by

$$h_{\max}(s, h, \ell) \equiv \min\{Q_a - \ell_a(s, h, \ell), Q_{a,\text{high}}\} + Q_t,$$

and

$$\ell_{\max}(s, h, \ell) \equiv \min\{Q_a - h_a(s, h, \ell), Q_{a,\text{low}}\} + Q_t.$$

The corresponding state space for high- and low-priority data calls is visualised in Figure 3.4, where the light area contains system states with empty high- and low-priority access queues, while in states contained in the medium and dark areas either one or both transfer queues are completely filled, respectively, and thus either one or both access queues are non-empty.

The channel capacity that is available for the data service is shared by all data calls in the transfer queues according to a Discriminatory Processor Sharing service discipline, which is a natural generalisation of the basic PS scheme deployed in the SVD model with a single data class. Within each priority class all data calls in transfer are treated equally. The relative amount of attention given to a low-priority data call compared to a high-priority data call is given by a predetermined parameter $\phi \in [0, 1]$. In case $\phi = 0$ the low-priority class becomes a best effort class, only receiving service capacity if the high-priority class cannot fully utilise the channel capacity. At the other extreme, if $\phi = 1$ both priority classes are treated equally. The channel assignment scheme is defined by:

$$\beta_{\text{high}}(s, h, \ell) \equiv \min\left\{\beta_{\text{GPRS}}^{\max}, (h_t(s, h, \ell) + \phi\ell_t(s, h, \ell))^{-1}(C_{\text{total}} - s)\right\},$$

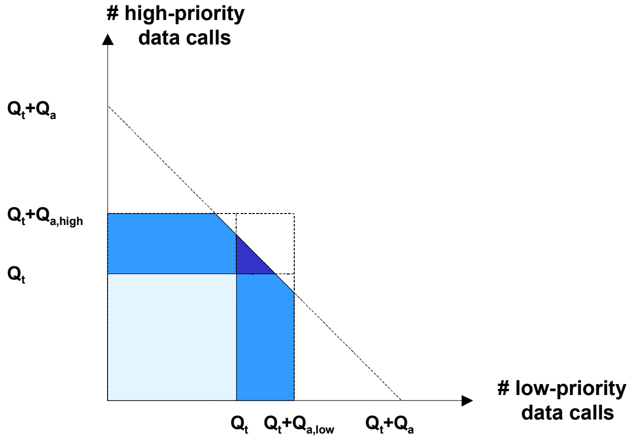

# high-priority data calls
Q_t+Q_a
Q_t+Q_{a,high}
Q_t
# low-priority data calls
Q_t
Q_t+Q_{a,low}
Q_t+Q_a

Figure 3.4 Illustration of the feasible state space for high- and low-priority data calls in the SHL model.

and

$$\beta_{\text{low}}(s,h,\ell) \equiv \min\left\{\beta_{\text{GPRS}}^{\max}, \ell_t(s,h,\ell)^{-1}\left((C_{\text{total}}-s)-h_t(s,h,\ell)\,\beta_{\text{high}}(s,h,\ell)\right)\right\},$$

where $C_{\text{total}}-s$ is the number of traffic channels available for data transfer in a system state given a presence of s speech calls in the system. By convention, we enforce that $\beta_{\text{high}}(s,h,\ell)=0$ ($\beta_{\text{low}}(s,h,\ell)=0$) if there are no active high-priority (low-priority) data calls, i.e. if $h=0$ ($\ell=0$).

If $\beta_{\text{GPRS}}^{\max}$ is not restrictive, the above channel assignment functions reduce to the more insightful expressions $\beta_{\text{high}}(s,h,\ell)=(h_t(s,h,\ell)+\phi\ell_t(s,h,\ell))^{-1}(C_{\text{total}}-s)$ and $\beta_{\text{low}}(s,h,\ell) = \phi\beta_{\text{high}}(s,h,\ell) = \phi(h_t(s,h,\ell)+\phi\ell_t(s,h,\ell))^{-1}(C_{\text{total}}-s)$, so that a low-priority data call receives indeed a fraction ϕ of the service capacity assigned to each high-priority data call.

The system state descriptors $h_t(s,h,\ell)$, $h_a(s,h,\ell)$, $\ell_t(s,h,\ell)$ and $\ell_a(s,h,\ell)$, the Call Admission Control thresholds $s_{\max}(s,h,\ell)$, $h_{\max}(s,h,\ell)$ and $\ell_{\max}(s,h,\ell)$, and the channel assignment functions $\beta_{\text{speech}}(s,h,\ell)$, $\beta_{\text{high}}(s,h,\ell)$ and $\beta_{\text{low}}(s,h,\ell)$, fully specify the studied channel sharing scheme and are thus the only ingredients required for our performance analysis. Note that seven parameters are to be set by the network operator, namely C_{speech}, C_{data}, Q_t, Q_a, $Q_{a,\text{high}}$, $Q_{a,\text{low}}$, and ϕ. All call handling procedures are then explicitly defined.

3.4.5. ON THE GENERALITY OF THE MODELS

A number of generalisations of the call handling schemes in the presented model can be made without complicating the performance analysis presented below, as long as model adjustments can be captured by the Call Admission Control thresholds and channel assignment schemes that have been introduced above. These generalisations have been consciously omitted here for clarity of presentation. Among the feasible generalisations we mention the following.

- The presented mathematical model is in principle suitable to analyse a system that integrates speech, video, high- and low-priority data calls in an SVHL model, as stated before. Although a broader range of GOS/QOS classes can be studied for the three service types, one is invariably limited by the curse of dimensionality, in the sense that the generation of numerical results may require solving excessively large systems of linear equations.

- Different channel sharing schemes. For the SVD and the SHL models, a single channel sharing scheme has been proposed that is characterised by parameters C_{speech}, C_{video} and C_{data}. Although this is beyond the scope of the present study, one can readily formulate different channel sharing schemes, e.g. a total sharing scheme without any reservations for specific service types. Such alternatives affect the Call Admission Control thresholds, the channel assignment functions and the system state descriptors (see also [151] and Section 2.7).

- Unequal transfer queue sizes for high- and low-priority data calls in the SHL model, i.e. $Q_{t,\text{high}}$ and $Q_{t,\text{low}}$. This generalisation is readily implemented by substituting Q_t by $Q_{t,\text{high}}$ or $Q_{t,\text{low}}$ in the Call Admission Control thresholds $h_{\max}(s,h,\ell)$ and $\ell_{\max}(s,h,\ell)$, respectively, as well as in the corresponding system state descriptors $h_t(s,h,\ell)$, $h_a(s,h,\ell)$, $\ell_t(s,h,\ell)$, and $\ell_a(s,h,\ell)$.

- (Priority class-specific) minimum transfer rates for active data calls to ensure some minimum QOS. Such a generalisation, e.g. $\beta_{\text{data}}^{\min}$ in the SVD model can be implemented either by reserving some capacity for data transfer ($C_{\text{data}} > 0$) and adjusting the transfer queue size Q_t such that $C_{\text{data}}/Q_t \geq \beta_{\text{data}}^{\min}$, or else by appropriately adjusting all Call Admission Control and channel assignment schemes to ensure that in any feasible state sufficient resources are available to award the minimum transfer rate to each of the present data calls.

- Service-specific maximum transfer rates for video and/or data calls, e.g. $\beta_{\text{low}}^{\max}$ in order to enforce more apparent QOS differentiation between high- and low-priority data calls (SHL model), or to differentiate between the technical limitations of video and data terminals. Such a generalisation is readily implemented by substituting a service specific $\beta_{\text{data}}^{\max}$ or $\beta_{\text{video}}^{\max}$ in $\beta_{\text{video}}(s, v, d)$, $\beta_{\text{data}}(s, v, d)$, $\beta_{\text{high}}(s, h, \ell)$, and/or $\beta_{\text{low}}(s, h, \ell)$. It is obvious that such maximum transfer rates must exceed any minimum transfer rates implemented, e.g. $\beta_{\text{video}}^{\max} \geq \beta_{\text{video}}^{\min}$.

- Restriction of video call assignments to be limited to a number of prefixed levels, e.g. 2, 4 or 8 traffic channels if this corresponds more accurately to an assumed scalable video coding algorithm (see e.g. [14]). Such a generalisation affects the video call channel assignment function $\beta_{\text{video}}(s, v, d)$, leading to a mathematically rather unappealing yet still tractable expression.

Besides these generalisations of the proposed call handling schemes, we note that with respect to call characteristics high- and low-priority data calls need not have the same exponential call length distribution: mean call sizes μ_{high} and μ_{low} can be specified separately. The corresponding model may then also be interpreted as differentiating based on the data call size class rather than the priority level. In the sensitivity analysis presented in Chapter 4 we will depart from the exponentiality assumption and assess the impact of the data call size variability, for models with and without data QOS differentiation.

3.4.6. PERFORMANCE MEASURES

The system level performance in each model is assessed in terms of the *expected channel utilisation*. The GOS of the speech, video and data service is expressed by the *call blocking probabilities*, while the experienced video and data QOS are primarily given by the *(conditional) expected video throughput* and the *(conditional) expected data call sojourn time*, respectively.

3.5. BASIC PERFORMANCE ANALYSIS

Now that both models are formulated, we are able to present our performance analysis. In this section, the system evolution of each model is formulated as a continuous-time Markov chain, and the equilibrium distribution is determined. Subsequently, some basic performance measures are given, that can be calculated directly from the equilibrium distribution of the Markov chains.

3.5.1. SPEECH/VIDEO/DATA MODEL

The evolution of the system in the SVD model can be described by an irreducible three-dimensional continuous-time *Markov chain* $(S(t), V(t), D(t))_{t\geq 0}$, with states denoted (s, v, d). The state space of the Markov chain is given by

$$\begin{aligned}\mathbb{S} \;\equiv\; & \Big\{(s,v,d) \in \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0 : \\ & s \leq s_{\max}(s,v,d) \text{ and } v \leq v_{\max}(s,v,d) \text{ and } d \leq d_{\max}(s,v,d)\Big\}.\end{aligned}$$

Ordering $\mathbb{S}$ lexicographically in (s, v, d), the infinitesimal generator is given by

$$\mathcal{Q} \equiv \begin{pmatrix} \mathcal{C}_0 & \mathcal{A}_0 & \mathcal{O} & \cdots & \mathcal{O} \\ \mathcal{B}_1 & \mathcal{C}_1 & \ddots & \ddots & \vdots \\ \mathcal{O} & \ddots & \ddots & \ddots & \mathcal{O} \\ \vdots & \ddots & \ddots & \mathcal{C}_{s_{\max}(0,0,0)-1} & \mathcal{A}_{s_{\max}(0,0,0)-1} \\ \mathcal{O} & \cdots & \mathcal{O} & \mathcal{B}_{s_{\max}(0,0,0)} & \mathcal{C}_{s_{\max}(0,0,0)} \end{pmatrix} \in \mathbb{R}^{|\mathbb{S}|} \times \mathbb{R}^{|\mathbb{S}|},$$

where $|\mathbb{S}| = \sum_{s=0}^{s_{\max}(0,0,0)} o_{\max}(s)$, with $o_{\max}(s) \equiv \sum_{v=0}^{v_{\max}(s,0,0)} (d_{\max}(s,v,0)+1)$ the number of system states with s speech calls. The blocks of $\mathcal{Q}$ are typically not square as $o_{\max}(s)$ is typically different from $o_{\max}(s+1)$.

The super-diagonal blocks $\mathcal{A}_s \equiv \lambda_{\text{speech}}\, \mathcal{I}_{o_{\max}(s)}$, $s = 0, \cdots, s_{\max}(0,0,0) - 1$, generate speech call arrival events, where $\mathcal{I}_{o_{\max}(s)}$ is the $o_{\max}(s) \times o_{\max}(s)$ identity matrix. The sub-diagonal blocks $\mathcal{B}_s \equiv s\, \mu_{\text{speech}}\, \mathcal{I}_{o_{\max}(s)}$, $s = 1, \cdots, s_{\max}(0,0,0)$, generate speech call termination events. Finally, the blocks $\mathcal{C}_s$, $s = 0, \cdots, s_{\max}(0,0,0)$, on the diagonal generate video and data call arrival and termination events, and have the following structure:

$$\mathcal{C}_s \equiv \begin{pmatrix} \mathcal{C}^c_{s,0} & \mathcal{C}^a_{s,0} & \mathcal{O} & \cdots & \mathcal{O} \\ \mathcal{C}^b_{s,1} & \mathcal{C}^c_{s,1} & \ddots & \ddots & \vdots \\ \mathcal{O} & \ddots & \ddots & \ddots & \mathcal{O} \\ \vdots & \ddots & \ddots & \mathcal{C}^c_{s,v_{\max}(s,0,0)-1} & \mathcal{C}^a_{s,v_{\max}(s,0,0)-1} \\ \mathcal{O} & \cdots & \mathcal{O} & \mathcal{C}^b_{s,v_{\max}(s,0,0)} & \mathcal{C}^c_{s,v_{\max}(s,0,0)} \end{pmatrix} \in \mathbb{R}^{o_{\max}(s)} \times \mathbb{R}^{o_{\max}(s)}.$$

The super-diagonal blocks $\mathcal{C}^a_{s,v}$, $v = 0, \cdots, v_{\max}(s,0,0) - 1$, generate video call arrival events, have dimensions $(d_{\max}(s,v,0)+1) \times (d_{\max}(s,v+1,0)+1) = (Q_a + Q_t \ +1) \times (Q_a + Q_t + 1)$, and entries $\mathcal{C}^a_{s,v}\left((s,v,d)\,;\ \ (s,v+1,d)\right) \equiv \lambda_{\text{video}}$, $d = 0, \cdots, d_{\max}(s,v,0)$. All other entries of $\mathcal{C}^a_{s,v}$ are 0. The sub-diagonal blocks $\mathcal{C}^b_{s,v}$, $v = 1, \cdots, v_{\max}(s,0,0)$, generate video call termination events, have dimensions $(d_{\max}(s,v,0)+1) \times (d_{\max}(s,v-1,0)+1) = (Q_a + Q_t + 1) \times (Q_a + Q_t + 1)$, and entries $\mathcal{C}^b_{s,v}\left((s,v,d)\,;(s,v-1,d)\right) \equiv v\,\mu_{\text{video}}$, $d = 0, \cdots, d_{\max}(s,v,0)$. All other entries of $\mathcal{C}^b_{s,v}$ are 0. Finally, the square blocks $\mathcal{C}^c_{s,v}$, $v = 0, \cdots, v_{\max}(s,0,0)$, on the diagonal generate data call arrival and termination events, have dimensions $(d_{\max}(s,v,0)+1) \times (d_{\max}(s,v,0)+1) = (Q_a + Q_t + 1) \times (Q_a + Q_t + 1)$, and entries $\mathcal{C}^c_{s,v}((s,v,d)\,;(s,v,d+1)) \equiv \lambda_{\text{data}}$, $d = 0, \cdots, d_{\max}(s,v,0) - 1$, and $\mathcal{C}^c_{s,v}\left((s,v,d)\,;(s,v,d-1)\right) \equiv \beta_{\text{data}}(s,v,d)\, d_t(s,v,d)\mu_{\text{data}}$, $d = 1, \cdots, d_{\max}(s,v,0)$. Furthermore, the diagonal entries of $\mathcal{C}^c_{s,v}$ are such that the entries of each row of $\mathcal{Q}$ sum up to 0. All other entries of $\mathcal{C}^c_{s,v}$ are equal to zero. Note that the super- and sub-diagonal blocks are square matrices since $d_{\max}(s,v,0) \equiv Q_a + Q_t$, for all supported s and v.

Since the finite state space Markov chain $(S(t), V(t), D(t))_{t \geq 0}$ is irreducible, a unique *equilibrium distribution* π exists that satisfies the system of global balance equations $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0}$, with $\mathbf{0}$ the vector with all entries zero and $\boldsymbol{\pi}$ lexicographically ordered in $(s,v,d) \in \mathbb{S}$.

A number of *basic performance measures* can be obtained directly from the equilibrium distribution of the considered Markov chain. From a system's perspective, the resource efficiency achieved can be measured by the *expected channel utilisation*,

$$\mathbf{U} \equiv C^{-1}_{\text{total}} \sum_{(s,v,d)\in\mathbb{S}} \pi(s,v,d)\ \left(s + \beta_{\text{video}}(s,v,d)\, v + \beta_{\text{data}}(s,v,d)\, d_t(s,v,d)\right).$$

The GOS of speech, video and data services is given by the *call blocking probabilities*,

$$\begin{aligned}
\mathbf{P}_{\text{speech}} &\equiv \sum_{(s,v,d)\in\mathbb{S}} \pi(s,v,d)1\left\{s = s_{\max}(s,v,d)\right\},\\
\mathbf{P}_{\text{video}} &\equiv \sum_{(s,v,d)\in\mathbb{S}} \pi(s,v,d)1\left\{v = v_{\max}(s,v,d)\right\},\\
\mathbf{P}_{\text{data}} &\equiv \sum_{(s,v,d)\in\mathbb{S}} \pi(s,v,d)1\left\{d = d_{\max}(s,v,d)\right\},
\end{aligned}$$

using the PASTA property [224].

The QOS delivered to the video service, expressed as the *expected (time-average) video throughput*, is the primary indicator of the user perceived audio/image quality. As the measure indicates the expected *per-call* video throughput, we must condition on the presence of at least one video call, obtaining

$$\mathbf{R}^t_{\text{video}} \equiv r_{\text{video}} \sum_{(s,v,d)\in\mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s,v,d)}{\sum_{(s,v,d)\in\mathbb{S}^+_{\text{video}}} \pi(s,v,d)} \right) \beta_{\text{video}}(s,v,d),$$

with $\mathbb{S}^+_{\text{video}} \equiv \{(s,v,d) \in \mathbb{S} : v > 0\}$ the set of states with at least one video call. Obviously, $r^{-1}_{\text{video}}\mathbf{R}^t_{\text{video}} \in \left[\beta^{\min}_{\text{video}}, \beta^{\max}_{\text{GPRS}}\right]$ must hold. Note that $\mathbf{R}^t_{\text{video}}$ is a *time*-average rather than a *call*-average throughput measure (hence the 't' superscript). In the next section a *call*-average video throughput measure (denoted $\mathbf{R}^c_{\text{video}}$) is derived which turns out to be different from the *time*-average throughput given here. Here we also refer to Chapter 5, which presents an extensive analysis and comparison of a range of throughput measures in Processor Sharing models.

For the data service, the delivered QOS, expressed as the *expected sojourn time* (*access time* plus *transfer time*) of a data call, is a performance measure of principal relevance to the user. In the SVD model, with the expected number of data calls in the access and transfer queues given by $\mathbf{N}_{a,\text{data}} \equiv \sum_{(s,v,d)\in\mathbb{S}} d_a(s,v,d)\,\pi(s,v,d)$ and $\mathbf{N}_{t,\text{data}} \equiv \sum_{(s,v,d)\in\mathbb{S}} d_t(s,v,d)\,\pi(s,v,d)$, respectively, Little's formula [224] is readily applied to determine the *expected access* and *transfer times* of a data call:

$$\mathbf{T}_{a,\text{data}} \equiv \frac{\mathbf{N}_{a,\text{data}}}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})} \text{ and } \mathbf{T}_{t,\text{data}} \equiv \frac{\mathbf{N}_{t,\text{data}}}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})},$$

respectively. The expected sojourn time of a data call in the SVD model is thus given by $\mathbf{T}_{\text{data}} \equiv \mathbf{T}_{a,\text{data}} + \mathbf{T}_{t,\text{data}}$. Another relevant performance measure characterising the Quality of Service delivered to data calls is the *expected (time-average) data throughput,* which for an *active* data call in the SVD model is given by

$$\mathbf{R}^t_{\text{data}} \equiv r_{\text{data}} \sum_{(s,v,d)\in\mathbb{S}^+_{\text{data}}} \left(\frac{\pi(s,v,d)}{\sum_{(s,v,d)\in\mathbb{S}^+_{\text{data}}} \pi(s,v,d)} \right) \beta_{\text{data}}(s,v,d),$$

with $\mathbb{S}^+_{\text{data}} \equiv \{(s,v,d) \in \mathbb{S} : d > 0\}$ the set of states with at least one (active) data call. Obviously, $r^{-1}_{\text{data}}\mathbf{R}^t_{\text{data}} \in (0, \beta^{\max}_{\text{GPRS}}]$ must hold. Note further that also $\mathbf{R}^t_{\text{data}}$ is a *time*-average rather than a *call*-average data throughput measure.

3.5.2. SPEECH/HIGH-/LOW-PRIORITY DATA MODEL

The evolution of the system in the SHL model can be described by an irreducible three-dimensional continuous-time *Markov chain* $(S(t), H(t), L(t))_{t\geq 0}$, with states denoted (s, h, ℓ). The state space of the Markov chain is given by

$$\begin{aligned} \mathbb{S} \quad \equiv \quad & \Big\{(s,h,\ell) \in \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0 : \\ & s \leq s_{\max}(s,h,\ell) \text{ and } h \leq h_{\max}(s,h,\ell) \text{ and } \ell \leq \ell_{\max}(s,h,\ell)\Big\}. \end{aligned}$$

Ordering $\mathbb{S}$ lexicographically in (v, h, ℓ), the infinitesimal generator is given by

$$\mathcal{Q} \equiv \begin{pmatrix} \mathcal{C}_0 & \mathcal{A}_0 & \mathcal{O} & \cdots & \mathcal{O} \\ \mathcal{B}_1 & \mathcal{C}_1 & \ddots & \ddots & \vdots \\ \mathcal{O} & \ddots & \ddots & \ddots & \mathcal{O} \\ \vdots & \ddots & \ddots & \mathcal{C}_{s_{\max}-1} & \mathcal{A}_{s_{\max}(0,0,0)-1} \\ \mathcal{O} & \cdots & \mathcal{O} & \mathcal{B}_{s_{\max}(0,0,0)} & \mathcal{C}_{s_{\max}(0,0,0)} \end{pmatrix} \in \mathbb{R}^{|\mathbb{S}|} \times \mathbb{R}^{|\mathbb{S}|},$$

with $|\mathbb{S}| = \sum_{s=0}^{s_{\max}(0,0,0)} o_{\max}(s)$, with $o_{\max}(s) \equiv \sum_{h=0}^{h_{\max}(s,0,0)} (\ell_{\max}(s,h,0)+1)$ the number of system states with s speech calls. As $o_{\max}(s)$ is equal for all s, all blocks in $\mathcal{Q}$ are square matrices of size $o_{\max}(s) \times o_{\max}(s)$.

The super-diagonal blocks $\mathcal{A}_s \equiv \lambda_{\text{speech}}\,\mathcal{I}_{o_{\max}(s)}$, $s = 0, \cdots, s_{\max}(0,0,0) - 1$, generate speech call arrival events. The sub-diagonal blocks $\mathcal{B}_s \equiv s\,\mu_{\text{speech}}\,\mathcal{I}_{o_{\max}(s)}$, $s = 1, \cdots, s_{\max}(0,0,0)$, generate speech call termination events. Finally, the blocks $\mathcal{C}_s$, $s = 0, \cdots, s_{\max}(0,0,0)$, on the diagonal generate data call arrival and termination events, and have the following structure:

$$\mathcal{C}_s \equiv \begin{pmatrix} \mathcal{C}^c_{s,0} & \mathcal{C}^a_{s,0} & \mathcal{O} & \cdots & \mathcal{O} \\ \mathcal{C}^b_{s,1} & \mathcal{C}^c_{s,1} & \ddots & \ddots & \vdots \\ \mathcal{O} & \ddots & \ddots & \ddots & \mathcal{O} \\ \vdots & \ddots & \ddots & \mathcal{C}^c_{s,h_{\max}(s,0,0)-1} & \mathcal{C}^a_{s,h_{\max}(s,0,0)-1} \\ \mathcal{O} & \cdots & \mathcal{O} & \mathcal{C}^b_{s,h_{\max}(s,0,0)} & \mathcal{C}^c_{s,h_{\max}(s,0,0)} \end{pmatrix} \in \mathbb{R}^{o_{\max}(s)} \times \mathbb{R}^{o_{\max}(s)}.$$

The super-diagonal blocks $\mathcal{C}^a_{s,h}$, $h = 0, \cdots, h_{\max}(s,0,0)-1$, generate high-priority data call arrival events, have dimensions $(\ell_{\max}(s,h,0)+1) \times (\ell_{\max}(s,h+1,0)+1)$, and entries $\mathcal{C}^a_{s,h}\left((s,h,\ell)\,;(s,h+1,\ell)\right) \equiv \xi\,\lambda_{\text{data}}$, $\ell = 0, \cdots, \ell_{\max}(s,h+1,0)$. All other entries of $\mathcal{C}^a_{s,h}$ are 0. The sub-diagonal blocks $\mathcal{C}^b_{s,h}$, $h = 1, \cdots, h_{\max}(s,0,0)$, generate high-priority data call termination events, have dimensions $(\ell_{\max}(s,h,0)+1) \times (\ell_{\max}(s,h-1,0)+1)$, and entries $\mathcal{C}^b_{s,h}\left((s,h,\ell)\,;(s,h-1,\ell)\right) \equiv \beta_{\text{high}}(s,h,\ell)\,h_t(s,h,\ell)\,\mu_{\text{data}}$, $\ell = 0, \cdots, \ell_{\max}(s,h,0)$. All other entries of $\mathcal{C}^b_{s,h}$ are 0. Finally, the square blocks $\mathcal{C}^c_{s,h}$, $h = 0, \cdots, h_{\max}(s,0,0)$, on the diagonal generate low-priority data call arrival and termination events, have dimensions $(\ell_{\max}(s,h,0)+1) \times (\ell_{\max}(s,h,0)+1)$, and entries $\mathcal{C}^c_{s,h}\left((s,h,\ell)\,;(s,h,\ell+1)\right) \equiv (1-\xi)\,\lambda_{\text{data}}$, $\ell = 0, \cdots, \ell_{\max}(s,h,0)-1$, and $\mathcal{C}^c_{s,h}\left((s,h,\ell)\,;(s,h,\ell-1)\right) \equiv \beta_{\text{low}}(s,h,\ell)\,\ell_t(s,h,\ell)\,\mu_{\text{data}}$, $\ell = 1, \cdots, \ell_{\max}(s,h,0)$. Furthermore, the diagonal entries of $\mathcal{C}^c_{s,h}$ are such that the entries of each row of $\mathcal{Q}$ sum up to 0. All other entries of $\mathcal{C}^c_{s,h}$ are equal to zero.

Since the finite state space Markov chain $(S(t), H(t), L(t))_{t\geq 0}$ is irreducible, a unique *equilibrium distribution* π exists that satisfies the system of global balance equations $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0}$, with $\mathbf{0}$ the vector with all entries zero and $\boldsymbol{\pi}$ lexicographically ordered in $(s,h,\ell) \in \mathbb{S}$.

Basic performance measures for the SHL model can be obtained analogously to those specified for the SVD model. Here we confine ourselves to listing the relevant measures: expected channel utilisation $\mathbf{U}$, blocking probabilities $\mathbf{P}_{\text{speech}}$, $\mathbf{P}_{\text{high}}$ and $\mathbf{P}_{\text{low}}$ for speech, high- and low-priority data calls respectively, QOS measures $\mathbf{R}^t_{\text{high}}$

and $\mathbf{R}^t_{\text{low}}$ denoting the expected (time-average) throughput of *active* high- and low-priority data calls, respectively; and the QOS measures $\mathbf{T}_{a,\text{high}}$, $\mathbf{T}_{t,\text{high}}$, $\mathbf{T}_{\text{high}}$, $\mathbf{T}_{a,\text{low}}$, $\mathbf{T}_{t,\text{low}}$, and $\mathbf{T}_{\text{low}}$, denoting the expected access, transfer and sojourn times of high- and low-priority data calls, respectively. In the SHL model the speech service is unaffected by the data service, hence the speech call blocking probability is readily determined using the Erlang loss probability (e.g. [213]), with load ρ_{speech} and number of channels $C_{\text{speech}} - C_{\text{data}}$.

3.6. CONDITIONAL PERFORMANCE ANALYSIS

The section presents a conditional performance analysis. More specifically, we derive expressions for the conditional QOS of an elastic video or data call of a given duration or size, respectively, and admitted to the system in a given state. This conditional analysis is considerably more involved than the derivation of the basic performance measures presented above, which are obtained immediately from the equilibrium distributions. As the analysis of the QOS of data calls is analogous to the analyses presented in Chapter 2 and in [174, 175], the data QOS performance expressions are stated without proofs.

3.6.1. SPEECH/VIDEO/DATA MODEL

For the SVD model, we determine the expected throughput of a video call conditional on its duration and the system state upon arrival, and the expected sojourn time of a data call conditional on its size and the system state upon arrival.

CONDITIONAL VIDEO QOS ANALYSIS

As demonstrated above, the expected throughput $\mathbf{R}^t_{\text{video}}$ of a video call is readily calculated from the equilibrium distribution $\pi(s,v,d)$, $(s,v,d) \in \mathbb{S}$. It was noted however that the obtained throughput measure is a *time*-average measure, whereas in this section a *call*-average throughput measure is determined, which is undeniably the most appropriate throughput measure from a call's perspective rather than from the system's perspective.

Recall from the previous section that $\mathbb{S}^+_{\text{video}} \equiv \{(s,v,d) \in \mathbb{S} : v > 0\}$ contains all system states with at least one video call present. For each state $(s,v,d) \in \mathbb{S}^+_{\text{video}}$

define $x_{s,v,d}(\tau)$ as the random number of kbits (*transfer volume*) transmitted by an admitted video call of duration τ, arriving at a given system state (s, v, d) with s speech calls, v video calls, and d data calls (v *includes* the new video call), and let $\widehat{x}_{s,v,d}(\tau) \equiv \mathbf{E}\{x_{s,v,d}(\tau)\}$ denote its expectation. Then the corresponding expected throughput is equal to $\widehat{x}_{s,v,d}(\tau)/\tau$, while the expected throughput of an *admitted* video call of duration τ is given by (in kbits/s)

$$\mathbf{R}^{c}_{\text{video}}(\tau) \equiv \sum_{(s,v,d)\in\mathbb{S}^{+}_{\text{video}}} \frac{\widehat{x}_{s,v,d}(\tau)}{\tau} \left(\frac{\pi(s, v-1, d)}{\sum\limits_{(s,v,d)\in\mathbb{S}^{+}_{\text{video}}} \pi(s, v-1, d)} \right), \tag{3.1}$$

where

$$\sum_{(s,v,d)\in\mathbb{S}^{+}_{\text{video}}} \pi(s, v-1, d) = \sum_{\substack{(s,v,d)\in\mathbb{S} \\ 0\leq v<v_{\max}(s,v,d)}} \pi(s, v, d) = 1 - \mathbf{P}_{\text{video}},$$

i.e. the term $\pi(s, v-1, d)/(1-\mathbf{P}_{\text{video}})$ in expression (3.1) is the equilibrium probability that the system is in state $(s, v-1, d)$, conditioned on the admission of an arriving video call, which is *not* equal to the probability $\pi(s, v-1, d)$ that an arbitrary video call finds the system in state $(s, v-1, d)$ upon arrival (PASTA), since the arrival process of *admitted* video calls is not Poisson. Integrating $\mathbf{R}^{c}_{\text{video}}(\tau)$ over the Probability Density Function (PDF) of τ yields the (unconditional) expected (call-average) throughput of an admitted video call:

$$\mathbf{R}^{c}_{\text{video}} \equiv \int_{\tau=0}^{\infty} \mathbf{R}^{c}_{\text{video}}(\tau)\mu_{\text{video}} \exp\{-\tau\mu_{\text{video}}\}\, d\tau,$$

which unfortunately turns out to be too difficult to be evaluate symbolically. We stress however, that in general the *time*-average video throughput $\mathbf{R}^{t}_{\text{video}}$ and the *call*-average video throughput $\mathbf{R}^{c}_{\text{video}}$ need not be the same.

In the following we derive an explicit expression for the vector $\widehat{\mathbf{x}}(\tau) = (\widehat{x}_{s,v,d}(\tau), (s, v, d) \in \mathbb{S}^{+}_{\text{video}})$. To this end, we need to make a modification to the original Markov chain, and hence introduce another infinitesimal generator, which is denoted $\mathcal{Q}^{\star}_{\text{video}}$. The modified chain is characterised by the presence of *one permanent video call*, i.e. there is one video call that never leaves the system, but shares in the available traffic channels as if it were an ordinary video call. This permanent video call is the tagged call whose throughput is to be determined, while the behaviour of all other calls is unchanged. Infinitesimal generator $\mathcal{Q}^{\star}_{\mathrm{video}}$ is similar to $\mathcal{Q}$, but of smaller dimensions, since the rows and columns corresponding to all states $(s, v, d) \notin \mathbb{S}^{+}_{\mathrm{video}}$ are crossed out. For all $(s, v, d) \in \mathbb{S}^{+}_{\mathrm{video}}$, the video call departure rates are modified as follows:

$$\mathcal{Q}^{\star}_{\mathrm{video}}\left((s, v, d)\,;(s, v-1, d)\right) = (v-1)\mu_{\mathrm{video}},$$

while the diagonal elements of $\mathcal{Q}^{\star}_{\mathrm{video}}$ are such that the entries of each row of $\mathcal{Q}^{\star}_{\mathrm{video}}$ sum up to 0. Let $\mathcal{B}_{\mathrm{video}} \equiv \mathit{diag}(\beta_{\mathrm{video}}(s, v, d),\ (s, v, d) \in \mathbb{S}^{+}_{\mathrm{video}})$ denote the diagonal matrix of average video channel assignments, lexicographically ordered in (s, v, d).

We first derive a closed-form expression for the Laplace-Stieltjes transform of the distribution of the conditional transfer volume $x_{s,v,d}(\tau)$. The obtained Laplace-Stieltjes transform is subsequently used to derive an expression for the conditional expected video throughput. An alternative, more direct derivation of the conditional expected video throughput can be carried out along the lines of the proof of Proposition 3.4 below. Define the Laplace-Stieltjes transform of the distribution of $x_{s,v,d}(\tau)$ by

$$X_{s,v,d}(\zeta, \tau) \equiv \mathbf{E}\left\{\exp\left\{-\zeta\, x_{s,v,d}(\tau)\right\}\right\},\ \ \mathrm{Re}(\zeta) \geq 0,\ \ (s, v, d) \in \mathbb{S}^{+}_{\mathrm{video}},$$

and let $\mathbf{X}(\zeta, \tau)$ be the vector with the $X_{s,v,d}(\zeta, \tau)$ ordered lexicographically in $(s, v, d) \in \mathbb{S}^{+}_{\mathrm{video}}$.

Lemma 3.1 *For $\tau \geq 0$ and $\mathrm{Re}(\zeta) \geq 0$, $\mathbf{X}(\zeta, \tau)$ satisfies the following differential equation and initial condition:*

$$\begin{aligned} \frac{\partial}{\partial \tau}\mathbf{X}(\zeta, \tau) &= \left(\mathcal{Q}^{\star}_{\mathit{video}} - \zeta r_{\mathit{video}}\mathcal{B}_{\mathit{video}}\right)\mathbf{X}(\zeta, \tau), & (3.2)\\ \mathbf{X}(\zeta, 0) &= \mathbf{1}, & (3.3) \end{aligned}$$

with the unique solution given by

$$\mathbf{X}(\zeta, \tau) = \exp\left\{\tau\left(\mathcal{Q}^{\star}_{\mathit{video}} - \zeta r_{\mathit{video}}\mathcal{B}_{\mathit{video}}\right)\right\}\mathbf{1}. \qquad (3.4)$$

Proof The first part of the lemma is proven by marginal analysis. Consider a time interval of length $\Delta > 0$, with Δ sufficiently small such that the tagged video call cannot terminate within this time, i.e. $\Delta < \tau$. Condition on all the possible events occurring in this interval, starting in state $(s, v, d) \in \mathbb{S}^{+}_{\text{video}}$. For notational convenience and readability, the boundary constraints are not explicitly considered. Equations for the boundary can be derived by analogy with the results below.

$$
\begin{aligned}
& X_{s,v,d}(\zeta, \tau) \\
&\quad \equiv \mathbf{E}\left\{\exp\left\{-\zeta\, x_{s,v,d}(\tau)\right\}\right\} \\
&\quad = \lambda_{\text{speech}} \Delta \exp[-\zeta(r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,(\Delta - O\,(\Delta)) \\
&\qquad\qquad + r_{\text{video}} \beta_{\text{video}}\,(s+1, v, d)\, O\,(\Delta))]\ X_{s+1,v,d}(\zeta, \tau - \Delta) \\
&\qquad + s\mu_{\text{speech}} \Delta \exp[-\zeta(r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,(\Delta - O\,(\Delta)) \\
&\qquad\qquad + r_{\text{video}} \beta_{\text{video}}\,(s-1, v, d)\, O\,(\Delta))]\ X_{s-1,v,d}(\zeta, \tau - \Delta) \\
&\qquad + \lambda_{\text{video}} \Delta \exp[-\zeta(r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,(\Delta - O\,(\Delta)) \\
&\qquad\qquad + r_{\text{video}} \beta_{\text{video}}\,(s, v+1, d)\, O\,(\Delta))]\ X_{s,v+1,d}(\zeta, \tau - \Delta) \\
&\qquad + (v-1)\,\mu_{\text{video}} \Delta \exp[-\zeta(r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,(\Delta - O\,(\Delta)) \\
&\qquad\qquad + r_{\text{video}} \beta_{\text{video}}\,(s, v-1, d)\, O\,(\Delta))]\ X_{s,v-1,d}(\zeta, \tau - \Delta) \\
&\qquad + \lambda_{\text{data}} \Delta \exp[-\zeta(r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,(\Delta - O\,(\Delta)) \\
&\qquad\qquad + r_{\text{video}} \beta_{\text{video}}\,(s, v, d+1)\, O\,(\Delta))]\ X_{s,v,d+1}(\zeta, \tau - \Delta) \\
&\qquad + \beta_{\text{data}}\,(s, v, d)\, d_t(s, v, d) \mu_{\text{data}} \Delta \exp[-\zeta(r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,(\Delta - O\,(\Delta)) \\
&\qquad\qquad + r_{\text{video}} \beta_{\text{video}}\,(s, v, d-1)\, O\,(\Delta))]\ X_{s,v,d-1}(\zeta, \tau - \Delta) \\
&\qquad + (-\lambda_{\text{speech}} \Delta - s\mu_{\text{speech}} \Delta - \lambda_{\text{video}} \Delta - (v-1)\,\mu_{\text{video}} \Delta - \lambda_{\text{data}} \Delta \\
&\qquad\qquad - \beta_{\text{data}}(s, v, d) d_t(s, v, d) \mu_{\text{data}} \Delta) \exp\left[-\zeta r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,\Delta\right] \\
&\qquad\qquad \times X_{s,v,d}(\zeta, \tau - \Delta) \\
&\qquad + \left(1 - \zeta r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,\Delta + \sum_{j=2}^{\infty} \frac{(-\zeta r_{\text{video}} \beta_{\text{video}}\,(s, v, d)\,\Delta)^j}{j!}\right) \\
&\qquad\qquad \times X_{s,v,d}(\zeta, \tau - \Delta) \\
&\qquad + o(\Delta).
\end{aligned}
$$

Rearranging terms yields

$$\begin{aligned}
&\frac{X_{s,v,d}(\zeta,\tau)-X_{s,v,d}(\zeta,\tau-\Delta)}{\Delta}\\
&\quad= \lambda_{\text{speech}}\exp[-\zeta(r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,(\Delta-O\,(\Delta))\\
&\qquad\qquad +r_{\text{video}}\beta_{\text{video}}\,(s+1,v,d)\,O\,(\Delta))]\;X_{s+1,v,d}(\zeta,\tau-\Delta)\\
&\qquad +s\mu_{\text{speech}}\exp[-\zeta(r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,(\Delta-O\,(\Delta))\\
&\qquad\qquad +r_{\text{video}}\beta_{\text{video}}\,(s-1,v,d)\,O\,(\Delta))]\,X_{s-1,v,d}(\zeta,\tau-\Delta)\\
&\qquad +\lambda_{\text{video}}\exp[-\zeta(r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,(\Delta-O\,(\Delta))\\
&\qquad\qquad +r_{\text{video}}\beta_{\text{video}}\,(s,v+1,d)\,O\,(\Delta))]\,X_{s,v+1,d}(\zeta,\tau-\Delta)\\
&\qquad +(v-1)\,\mu_{\text{video}}\exp[-\zeta(r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,(\Delta-O\,(\Delta))\\
&\qquad\qquad +r_{\text{video}}\beta_{\text{video}}\,(s,v-1,d)\,O\,(\Delta))]\,X_{s,v-1,d}(\zeta,\tau-\Delta)\\
&\qquad +\lambda_{\text{data}}\exp[-\zeta(r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,(\Delta-O\,(\Delta))\\
&\qquad\qquad +r_{\text{video}}\beta_{\text{video}}\,(s,v,d+1)\,O\,(\Delta))]\,X_{s,v,d+1}(\zeta,\tau-\Delta)\\
&\qquad +\beta_{\text{data}}\,(s,v,d)\,d_t(s,v,d)\mu_{\text{data}}\exp[-\zeta(r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,(\Delta-O\,(\Delta))\\
&\qquad\qquad +r_{\text{video}}\beta_{\text{video}}\,(s,v,d-1)\,O\,(\Delta))]\;X_{s,v,d-1}(\zeta,\tau-\Delta)\\
&\qquad +(-\lambda_{\text{speech}}-s\mu_{\text{speech}}-\lambda_{\text{video}}-(v-1)\,\mu_{\text{video}}-\lambda_{\text{data}}\\
&\qquad\qquad -\beta_{\text{data}}(s,v,d)d_t(s,v,d)\,\mu_{\text{data}})\exp\left[-\zeta r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,\Delta\right]\\
&\qquad\qquad \times X_{s,v,d}(\zeta,\tau-\Delta)\\
&\qquad +\left(-\zeta r_{\text{video}}\beta_{\text{video}}\,(s,v,d)+\frac{1}{\Delta}\sum_{j=2}^{\infty}\frac{(-\zeta r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,\Delta)^j}{j!}\right)\\
&\qquad\qquad \times X_{s,v,d}(\zeta,\tau-\Delta)\\
&\qquad +\frac{o(\Delta)}{\Delta},
\end{aligned}$$

and letting $\Delta\downarrow 0$ gives the system of differential equations

$$\begin{aligned}
\frac{\partial X_{s,v,d}(\zeta,\tau)}{\partial\tau} &= \lambda_{\text{speech}}X_{s+1,v,d}(\zeta,\tau)+s\mu_{\text{speech}}X_{s-1,v,d}(\zeta,\tau)\\
&\quad +\lambda_{\text{video}}X_{s,v+1,d}(\zeta,\tau)+(v-1)\,\mu_{\text{video}}\;X_{s,v-1,d}(\zeta,\tau)\\
&\quad +\lambda_{\text{data}}X_{s,v,d+1}(\zeta,\tau)+\beta_{\text{data}}\,(s,v,d)\,d_t(s,v,d)\mu_{\text{data}}X_{s,v,d-1}(\zeta,\tau)
\end{aligned}$$

$$
\begin{aligned}
&+(-\lambda_{\text{speech}} - s\mu_{\text{speech}} - \lambda_{\text{video}} - (v-1)\,\mu_{\text{video}} - \lambda_{\text{data}} \\
&\qquad -\beta_{\text{data}}(s,v,d)d_t(s,v,d)\mu_{\text{data}})X_{s,v,d}(\zeta,\tau) \\
&-\zeta r_{\text{video}}\beta_{\text{video}}\,(s,v,d)\,X_{s,v,d}(\zeta,\tau),
\end{aligned}
$$

using the continuity of $X_{s,v,d}(\zeta,\tau)$ in τ. The system of differential equations may equivalently be written in the matrix notation of expression (3.2). The initial condition (3.3) simply reflects the fact that the transfer volume $x_{s,v,d}(0)$ of a video call with a duration of zero seconds equals zero bits:

$$\mathbf{X}(\zeta,0) = \left(\mathbf{E}\left\{\exp\left\{-\zeta x_{s,v,d}(0)\right\}\right\}\right)_{(s,v,d)\in\mathbb{S}^{+}_{\text{video}}} = \mathbf{1}.$$

The *existence* and *uniqueness* of a solution $\mathbf{X}(\zeta,\tau)$ to the system of differential equations (3.2) with initial condition (3.3) follows from e.g. [51, Chapter 1, Section 8]. To conclude the proof, it is readily verified that the claimed solution (3.4) indeed satisfies the system of differential equations (3.2) with initial condition (3.3). □

Using the closed-form expression (3.4) for the Laplace-Stieltjes transform of the distribution of $x_{s,v,d}(\tau)$, an explicit expression of the conditional expected transfer volume and, consequently, the conditional expected video throughput can be derived.

Proposition 3.1 *Let $\boldsymbol{\pi}^{\star}_{video} \equiv \left(\pi^{\star}_{video}(s,v,d),\ (s,v,d)\in\mathbb{S}^{+}_{video}\right)$ be the equilibrium probability distribution vector corresponding to the Markov chain with one permanent video call, i.e. $\boldsymbol{\pi}^{\star}_{video}\mathcal{Q}^{\star}_{video} = \mathbf{0}$. Further, let $\boldsymbol{\gamma}_{video} \equiv \left(\gamma_{video}(s,v,d),\ (s,v,d)\in\mathbb{S}^{+}_{video}\right)$ be the unique solution to*

$$
\begin{aligned}
\mathcal{Q}^{\star}_{video}\boldsymbol{\gamma}_{video} &= r_{video}\left(\left(\boldsymbol{\pi}^{\star}_{video}\mathcal{B}_{video}\mathbf{1}\right)\mathbf{1} - \mathcal{B}_{video}\mathbf{1}\right), && (3.5)\\
\boldsymbol{\pi}^{\star}\boldsymbol{\gamma}_{video} &= 0. && (3.6)
\end{aligned}
$$

Then the conditional expected transfer volumes $\widehat{\mathbf{x}}(\tau)$ are given by

$$\widehat{\mathbf{x}}(\tau) = \tau r_{video}\left(\boldsymbol{\pi}^{\star}_{video}\mathcal{B}_{video}\mathbf{1}\right)\mathbf{1} + \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{video}\right\}\right]\boldsymbol{\gamma}_{video}, \qquad (3.7)$$

so that the conditional expected throughput vector is given by

$$\frac{\widehat{\mathbf{x}}(\tau)}{\tau} = r_{video}\left(\boldsymbol{\pi}^{\star}_{video}\mathcal{B}_{video}\mathbf{1}\right)\mathbf{1} + \frac{1}{\tau}\left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{video}\right\}\right]\boldsymbol{\gamma}_{video}.$$

Proof Analogously to the proof of Proposition 2.4, the existence of a vector $\boldsymbol{\gamma}_{\text{video}}$ that satisfies (3.5) and its uniqueness up to a translation along the vector $\mathbf{1}$, are guaranteed by results in Markov reward chain theory. Interpreting $\boldsymbol{\gamma}_{\text{video}}$ as the vector of relative rewards in a Markov reward chain governed by the infinitesimal generator $\mathcal{Q}^{\star}_{\text{video}}$ and with immediate reward vector $\frac{1}{\eta} r_{\text{video}} \left(\mathcal{B}_{\text{video}}\mathbf{1} - \left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right)$ with η is the maximum rate of change in the Markov chain, and understanding that the long-term average rewards are zero, $\frac{1}{\eta}\boldsymbol{\pi}^{\star}_{\text{video}} r_{\text{video}} \left(\mathcal{B}_{\text{video}}\mathbf{1} - \left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right) = 0$, [213, Theorem 3.1, page 167] can be applied after a uniformisation of the continuous-time Markov chain. A translation of $\boldsymbol{\gamma}_{\text{video}}$ along the vector $\mathbf{1}$ does indeed not alter the solution, since for any $\alpha \in \mathbb{R}$, $\mathcal{Q}^{\star}_{\text{video}}\left(\boldsymbol{\gamma}_{\text{video}} + \alpha\mathbf{1}\right) = \mathcal{Q}^{\star}_{\text{video}}\boldsymbol{\gamma}_{\text{video}}$, or equivalently, $\left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\left(\boldsymbol{\gamma}_{\text{video}} + \alpha\mathbf{1}\right) = \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}}$, which readily follows from using the Taylor expansion of $\exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}$. The single degree of freedom that exists in choosing $\boldsymbol{\gamma}_{\text{video}}$ in expression (3.5), is used to normalise $\boldsymbol{\gamma}_{\text{video}}$ as in (3.6).

The vector of conditional expected transfer volumes $\widehat{\mathbf{x}}(\tau)$ is then obtained by taking the derivative of $\mathbf{X}(\zeta,\tau)$ with respect to ζ, and subsequently setting $\zeta = 0$.

$$\begin{aligned}
\widehat{\mathbf{x}}(\tau) &= -\frac{\partial}{\partial\zeta}\mathbf{X}(\zeta,\tau)\bigg|_{\zeta=0} \\
&= -\frac{\partial}{\partial\zeta}\sum_{k=0}^{\infty}\frac{\left(\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right) + \left(-\zeta\tau r_{\text{video}}\mathcal{B}_{\text{video}}\right)\right)^{k}}{k!}\mathbf{1}\Bigg|_{\zeta=0} \\
&= -\left(\sum_{k=1}^{\infty}\sum_{i=0}^{k-1}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-i-1}\left(-\tau r_{\text{video}}\mathcal{B}_{\text{video}}\right)\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{i}}{k!}\right)\mathbf{1} \\
&= \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\tau r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}} r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right) \\
&\qquad \times\left[\tau r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} - \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}} r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right] \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}} r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} - \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\tau\mathcal{Q}^{\star}_{\text{video}}\boldsymbol{\gamma}_{\text{video}} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}} r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left(\mathcal{I} - \sum_{k=0}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k}}{k!}\right)\boldsymbol{\gamma}_{\text{video}} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}} r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}}
\end{aligned}$$

where after the third equality sign only those matrix cross-products appear that remain after differentiating the terms in the preceding expression, and setting ζ to 0. The subsequent equality sign uses $\mathcal{Q}^{\star}_{\text{video}}\mathbf{1} = 0$, so that all terms with $i > 0$ disappear. A similar argument is used to obtain the fifth equality. Equation (3.5) is used for the sixth equality. □

Corollary 3.2 *The following asymptotic expressions immediately follow from the expressions in Proposition 3.1:*

$$\lim_{\tau\to\infty}\left\{\widehat{\mathbf{x}}(\tau) - \tau r_{video}\left(\boldsymbol{\pi}^{\star}_{video}\mathcal{B}_{video}\mathbf{1}\right)\mathbf{1}\right\} = \boldsymbol{\gamma}_{video},$$

and

$$\lim_{\tau\to\infty}\frac{\widehat{\mathbf{x}}(\tau)}{\tau} = r_{video}\left(\boldsymbol{\pi}^{\star}_{video}\mathcal{B}_{video}\mathbf{1}\right)\mathbf{1},$$

hence asymptotically the expected (call-average) video throughput $\widehat{x}_{s,v,d}(\tau)/\tau$ of a video call arriving in state (s,v,d) is equal to the expected (time-average) video throughput $r_{video}\boldsymbol{\pi}^{\star}_{video}\mathcal{B}_{video}\mathbf{1}$ in a system with one permanent video call, for all $(s,v,d) \in \mathbb{S}^{+}_{video}$.

Proof Since $\mathcal{Q}^{\star}_{\text{video}}$ is the infinitesimal generator of an irreducible finite state space Markov chain with equilibrium distribution vector $\boldsymbol{\pi}^{\star}_{\text{video}}$, $\lim_{\tau\to\infty}\exp\{\tau\mathcal{Q}^{\star}_{\text{video}}\} = \mathbf{1}\,\boldsymbol{\pi}^{\star}_{\text{video}}$, and thus $\lim_{\tau\to\infty}[\mathcal{I} - \exp\{\tau\mathcal{Q}^{\star}_{\text{video}}\}]\boldsymbol{\gamma}_{\text{video}} = \boldsymbol{\gamma}_{\text{video}}$, using (3.6), while $\lim_{\tau\to\infty}\tau^{-1}[\mathcal{I} - \exp\{\tau\mathcal{Q}^{\star}_{\text{video}}\}]\,\boldsymbol{\gamma}_{\text{video}} = \mathbf{0}$. □

Remark 3.1 Note that the asymptotic expression for the conditional expected (call-average) video throughput is given by $\widehat{\mathbf{x}}(\tau)/\tau = r_{\text{video}}\left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \boldsymbol{\gamma}_{\text{video}}/\tau$ which is non-linear in τ, as will also be illustrated in Section 3.7.

Remark 3.2 The asymptotic result is readily supported by an intuitive argument. Consider a video call of duration τ. As $\tau \to \infty$, the average assigned bit rate over the call's lifetime becomes deterministic (independent of τ) and is given by $r_{\text{video}}\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}$. Hence for very large τ the expected transfer volume is approximately equal to the deterministic call duration times the (almost) deterministic assigned bit rate, while the significance of the constant $\boldsymbol{\gamma}$ becomes negligible.

CONDITIONAL DATA QOS ANALYSIS

As demonstrated in Section 3.5, the expected sojourn time $\mathbf{T}_{\text{data}}$ of a data call is readily calculated from the equilibrium distribution $\pi(s,v,d)$, $(s,v,d) \in \mathbb{S}$. We now determine $\mathbf{T}_{\text{data}}(x)$, the conditional expected sojourn time of an admitted data call of size x, which is considerably more complicated. Since the conditional analysis is analogous to that presented in Chapter 2, we only state the main results here for completeness.

Recall from the previous section that $\mathbb{S}^{+}_{\text{data}} \equiv \{(s,v,d) \in \mathbb{S} : d > 0\}$ contains all system states with at least one data call present. Recall that if $(s,v,d) \in \mathbb{S}^{+}_{\text{data}}$ then $d_t(s,v,d) > 0$. For each state $(s,v,d) \in \mathbb{S}^{+}_{\text{data}}$ define $\sigma_{s,v,d}(\tau)$ as the random sojourn time of an admitted data call of size x, arriving at a given system state (s,v,d) with s speech calls, v video calls, and d data calls (d *includes* the new data call), and let $\widehat{\sigma}_{s,v,d}(x) \equiv \mathbf{E}\{\sigma_{s,v,d}(x)\}$ denote its expectation. Then the expected sojourn time of an *admitted* data call of size x is given by

$$\mathbf{T}_{\text{data}}(x) \equiv \frac{1}{1-\mathbf{P}_{\text{data}}} \sum_{(s,v,d)\in\mathbb{S}^{+}_{\text{data}}} \pi(s,v,d-1)\widehat{\sigma}_{s,v,d}(x). \tag{3.8}$$

Integrating $\widehat{\sigma}_{s,v,d}(x)$ over the PDF of x yields the conditional expected sojourn time of an admitted data call that arrives at system state (s,v,d). Unfortunately, the integral turns out to be too difficult to evaluate symbolically. Regarding the integral of $\mathbf{T}_{\text{data}}(x)$ over all possible values of τ, yielding the (unconditional) expected (call-average) throughput $\mathbf{T}_{\text{data}}$ it holds that

$$\mathbf{T}_{\text{data}} = \int\limits_{x=0}^{\infty} \mathbf{T}_{\text{data}}(x)\mu_{\text{data}} \exp\{-x\mu_{\text{data}}\}\, dx.$$

In the following an explicit expression for the vector $\widehat{\boldsymbol{\sigma}}(x) = (\widehat{\sigma}_{s,v,d}(x), (s,v,d) \in \mathbb{S}^{+}_{\text{data}})$, $x \in \mathbb{R}^{+}$, is derived, consisting of an access time and a transfer time component. By analogy with the sojourn time variables, let $\alpha_{s,v,d}$ denote the random access time of a newly admitted data call of size x, entering the system in state (s,v,d), which is obviously independent of the call size due to the access queue's FIFO discipline, and let $\widehat{\alpha}_{s,v,d} \equiv \mathbf{E}\{\alpha_{s,v,d}\}$ be its expectation. Similarly, denote with $\tau_{s_o,v_o,d_o}(x)$ the random transfer time of an admitted data call of size x, that starts its transfer in system state $(s_o, v_o, d_o) \in \mathbb{S}^+_{\text{data}}$, and let $\widehat{\tau}_{s_o,v_o,d_o}(x) \equiv \mathbf{E}\{\tau_{s_o,v_o,d_o}(x)\}$. Lastly, $\Psi_{\text{data}}\left((s,v,d)\,;(s_o,v_o,d_o)\right)$ is defined as the probability that a data call entering the system in state $(s,v,d) \in \mathbb{S}^+_{\text{data}}$ starts its transfer in state $(s_o,v_o,d_o) \in \mathbb{S}^+_{\text{data}}$.

Proposition 3.2 *The conditional expected sojourn time* $\widehat{\sigma}_{s,v,d}(x)$, $(s,v,d) \in \mathbb{S}^+_{\text{data}}$, $x \in \mathbb{R}^+$, *can be expressed as the sum of the expected access time and the expected transfer time, as follows:*

$$\widehat{\sigma}_{s,v,d}(x) = \widehat{\alpha}_{s,v,d} + \sum_{(s_o,v_o,d_o)\in\mathbb{S}^+_{\text{data}}} \widehat{\tau}_{s_o,v_o,d_o}(x)\,\Psi_{\text{data}}\left((s,v,d)\,;(s_o,v_o,d_o)\right). \tag{3.9}$$

Proof The result immediately follows from conditioning on the system state (s_o, v_o, d_o) where the tagged data call starts its transfer. □

Although an explicit closed-form expression can be found for the expected access times $\widehat{\alpha}_{s,v,d}$, $(s,v,d) \in \mathbb{S}^+_{\text{data}}$, along the lines developed in Section 2.6.1, these are not required to obtain $\mathbf{T}_{\text{data}}(x)$, which is readily verified by substitution of (3.9) in (3.8).

Corollary 3.3 *The conditional expected sojourn time of a data call* $\mathbf{T}_{\text{data}}(x)$, $x \in \mathbb{R}^+$, *can be calculated as follows:*

$$\begin{aligned}\mathbf{T}_{data}(x) \quad &= \quad \mathbf{T}_{a,data} + \\ \sum_{(s,v,d)\in\mathbb{S}^+_{\text{data}}} \sum_{(s_o,v_o,d_o)\in\mathbb{S}^+_{\text{data}}} &\widehat{\tau}_{s_o,v_o,d_o}(x)\Psi_{data}\left((s,v,d)\,;(s_o,v_o,d_o)\right)\frac{\pi(s,v,d-1)}{1-\mathbf{P}_{data}},\end{aligned}$$

where $\mathbf{T}_{a,data}$ *can be determined using the Markov chain's equilibrium distribution.*

In order to compute the *transition probability* matrix Ψ_{data} containing the probabilities $\Psi_{\text{data}}\left((s,v,d)\,;(s_o,v_o,d_o)\right)$ that a data call entering the system in state $(s,v,d) \in \mathbb{S}^+_{\text{data}}$ starts its transfer in state $(s_o,v_o,d_o) \in \mathbb{S}^+_{\text{data}}$, the state space $\mathbb{S}$ is augmented with an extra dimension. Denote with $N(t)$ the location in the queue of a tagged data call at time t, and let $N(t) = 0$ if this data call is no longer queued. The augmented Markov chain is denoted $(N(t), S(t), V(t), D(t))_{t\geq 0}$, with states (n,s,v,d) and state space

$$\mathbb{S}^\bullet \equiv \{(n,s,v,d) \in \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0 : n \leq d_t(s,v,d) \text{ and } (s,v,d) \in \mathbb{S}\}\,.$$

The state space is partitioned into the absorbing subset $\mathbb{S}_0^\bullet \equiv \{(n,s,v,d) \in \mathbb{S}^\bullet : n = 0\}$ and its transient complement, $\mathbb{S}_+^\bullet \equiv \mathbb{S}^\bullet \backslash \mathbb{S}_0^\bullet$. The augmented Markov chain is further modified by enforcing each state in $\mathbb{S}_0^\bullet$ to be absorbing. The state space $\mathbb{S}^\bullet$ of the modified Markov chain consists of $|\mathbb{S}_0^\bullet| = |\mathbb{S}|$ absorbing states and 1 transient class $\mathbb{S}_+^\bullet$.

The one-step transition probability matrix $\mathcal{P}_{\text{data}}^\bullet$ of the embedded jump chain associated with the augmented Markov chain can be written in the form

$$\mathcal{P}_{\text{data}}^\bullet \equiv \begin{pmatrix} \mathcal{I} & \mathcal{O} \\ \mathcal{P}_{\text{data},+0}^\bullet & \mathcal{P}_{\text{data},++}^\bullet \end{pmatrix},$$

after lexicographically ordering the states in $\mathbb{S}_0^\bullet$ and $\mathbb{S}_+^\bullet$, respectively. Here $\mathcal{I}$ is the identity matrix, $\mathcal{O}$ is the null-matrix, and $\mathcal{P}_{\text{data},+0}^\bullet$ and $\mathcal{P}_{\text{data},++}^\bullet$ are substochastic submatrices of $\mathcal{P}_{\text{data}}^\bullet$ corresponding to the transitions from $\mathbb{S}_+^\bullet$ to $\mathbb{S}_0^\bullet$, and from $\mathbb{S}_+^\bullet$ to $\mathbb{S}_+^\bullet$, respectively.

Proposition 3.3 *The probability* $\Psi_{data}\left((s,v,d)\,;(s_o,v_o,d_o)\right)$ *equals the element* $\Psi_{data}^\bullet$ $\left((d_a(s,v,d),s,v,d)\,;(0,s_o,v_o,d_o)\right)$ *of the probability matrix* $\Psi_{\text{data}}^\bullet$ *given by*

$$\Psi_{data}^\bullet = \left(\mathcal{I} - \mathcal{P}_{data,++}^\bullet\right)^{-1} \mathcal{P}_{data,+0}^\bullet.$$

Matrix $\Psi_{\text{data}}^\bullet$ *is of dimension* $|\mathbb{S}_+^\bullet| \times |\mathbb{S}_0^\bullet|$ *and contains the probabilities that the augmented chain, starting in any state in* $\mathbb{S}_+^\bullet$ *is eventually absorbed in a state in* $\mathbb{S}_0^\bullet$.

Proof The proof is analogous to that of Proposition 2.3 and follows from conditioning on the first transition out of (n,s,v,d). □

We now focus on the conditional expected *transfer time* $\widehat{\tau}_{s,v,d}(x)$ of a tagged active data call of length $x \geq 0$, starting its transfer in the presence of s speech calls, v video, and d data calls, $(s,v,d) \in \mathbb{S}_{\text{data}}^+$, including itself and all queued data calls. An explicit expression for the vector $\widehat{\boldsymbol{\tau}}(x) = \left(\widehat{\tau}_{s,v,d}(x),\ (s,v,d) \in \mathbb{S}_{\text{data}}^+\right)$ is derived below. The transfer time analysis requires another modification to the original Markov chain, which imposes the presence of *one permanently active data call*. Infinitesimal generator $\mathcal{Q}_{\text{data}}^\star$ of the modified Markov chain is similar to $\mathcal{Q}$, but of smaller dimensions, since the rows and columns corresponding to all states $(s,v,d) \notin \mathbb{S}_{\text{data}}^+$ are crossed out. The presence of the permanent data call is enforced by modifying the data call departure rates as follows:

$$\mathcal{Q}^{\star}_{\text{data}}\left((s,v,d)\,;(s,v,d-1)\right)=\beta_{\text{data}}(s,v,d)\left(d_t(s,v,d)-1\right)\mu_{\text{data}},$$

for all $(s,v,d)\in\mathbb{S}^{+}_{\text{data}}$, while the diagonal elements of $\mathcal{Q}^{\star}_{\text{data}}$ are modified accordingly. Let $\boldsymbol{\pi}^{\star}_{\text{data}}\equiv\left(\pi^{\star}_{\text{data}}(s,v,d),\ (s,v,d)\in\mathbb{S}^{+}_{\text{data}}\right)$ be the equilibrium distribution vector that satisfies the global balance equations of the modified Markov chain. Further, let $\mathcal{B}_{\text{data}}\equiv diag(\beta_{\text{data}}(s,v,d),\ (s,v,d)\in\mathbb{S}^{+}_{\text{data}})$ denote the lexicographically ordered diagonal matrix of average data transfer rates.

In contrast to the conditional data QOS analysis for HSCSD calls in Chapter 2, which were guaranteed a minimum assignment of one traffic channel, it is now possible that occasionally no resources are available for data transfer, i.e. $\beta_{\text{data}}\,(s,v,d)=0$ for some states $(s,v,d)\in\mathbb{S}^{+}_{\text{data}}$, depending on the model parameters. In the SVD model this only occurs if $C_{\text{data}}=0$, in which case the data calls are deprived from any capacity if $s+v\beta^{\max}_{\text{GPRS}}\geq C_{\text{total}}$. We therefore partition $\mathbb{S}^{+}_{\text{data}}$ into $\mathbb{S}^{+}_{\text{data},0}\equiv\left\{(s,v,d)\in\mathbb{S}^{+}_{\text{data}}:\beta_{\text{data}}\,(s,v,d)=0\right\}$ and its complement $\mathbb{S}^{+}_{\text{data},+}\equiv\mathbb{S}^{+}_{\text{data}}\backslash\mathbb{S}^{+}_{\text{data},0}$, and confine ourselves to a conditional analysis of $\widehat{\boldsymbol{\tau}}(x)$ for the more complex case that $\mathbb{S}^{+}_{\text{data},0}\neq\emptyset$. The alternate, simpler case can be analysed similarly.

Proposition 3.4 *Consider the case that $\mathbb{S}^{+}_{data,0}\neq\emptyset$. Partition*

$$\mathcal{Q}^{\star}_{data}=\begin{bmatrix}\mathcal{Q}^{\star}_{++} & \mathcal{Q}^{\star}_{+0}\\ \mathcal{Q}^{\star}_{0+} & \mathcal{Q}^{\star}_{00}\end{bmatrix},\ \mathcal{B}_{data}=\begin{bmatrix}\mathcal{B}_{+} & \mathcal{O}\\ \mathcal{O} & \mathcal{O}\end{bmatrix},\ \boldsymbol{\pi}^{\star}_{data}=\left(\boldsymbol{\pi}^{\star}_{data,0},\boldsymbol{\pi}^{\star}_{data,+}\right)$$

in accordance with the introduced state space partitioning, where we omit the 'data' subscript in the submatrices of $\mathcal{Q}^{\star}_{data}$ and $\mathcal{B}_{data}$ for enhanced readibility. Let $\boldsymbol{\gamma}_{data}\equiv\left(\gamma_{data}(s,v,d),\ (s,v,d)\in\mathbb{S}^{+}_{data,+}\right)$ be the unique solution to the system of linear equations

$$\begin{aligned}\left(\mathcal{Q}^{\star}_{++}+\mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\mathcal{Q}^{\star}_{0+}\right)\boldsymbol{\gamma}_{data} &= \frac{\mathcal{B}_{+}\mathbf{1}}{\boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\mathbf{1}}+ \qquad (3.10)\\ &\qquad -\left(\mathcal{I}+\mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\right)\mathbf{1},\\ \boldsymbol{\pi}^{\star}_{data,+}\mathcal{B}_{+}\boldsymbol{\gamma}_{data} &= 1. \qquad (3.11)\end{aligned}$$

Then the solution for $\widehat{\boldsymbol{\tau}}(x) = (\widehat{\boldsymbol{\tau}}_0(x), \widehat{\boldsymbol{\tau}}_+(x))$ *is given by*

$$\begin{cases} \widehat{\boldsymbol{\tau}}_0(x) = \left(-\mathcal{Q}^\star_{00}\right)^{-1} \left\{\mathbf{1} + \mathcal{Q}^\star_{0+}\widehat{\boldsymbol{\tau}}_+(x)\right\}, \\ \\ \widehat{\boldsymbol{\tau}}_+(x) = \dfrac{x}{\boldsymbol{\pi}^\star_{data,+}\mathcal{B}_+\mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{x\mathcal{B}_+^{-1}\left(\mathcal{Q}^\star_{++} + \mathcal{Q}^\star_{+0}\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}\right)\right\}\right]\boldsymbol{\gamma}_{data}, \end{cases} \tag{3.12}$$

while the following asymptotic expressions immediately follow:

$$\begin{cases} \lim_{x\to\infty}\left\{\widehat{\boldsymbol{\tau}}_0(x) - \dfrac{x}{\boldsymbol{\pi}^\star_{data,+}\mathcal{B}_+\mathbf{1}}\mathbf{1}\right\} = \left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathbf{1} + \left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}\boldsymbol{\gamma}_{data}, \\ \\ \lim_{x\to\infty}\left\{\widehat{\boldsymbol{\tau}}_+(x) - \dfrac{x}{\boldsymbol{\pi}^\star_{data,+}\mathcal{B}_+\mathbf{1}}\mathbf{1}\right\} = \boldsymbol{\gamma}_{data}, \end{cases} \tag{3.13}$$

indicating that for large data calls the expected sojourn time is approximately linear in the call size.

Proof In contrast to the conditional analysis of the video QOS, which first derived a closed-form expression for the Laplace-Stieltjes transform of the distribution of $x_{s,v,d}(t)$, a *direct* proof of Proposition 3.4 is given below, which is analogous to that of [174, Corollary 5.2].

We first concentrate on the evolution from state $(s, v, d) \in \mathbb{S}^+_{\text{data},0}$, and condition on the possible transitions:

$$\widehat{\tau}_{s,v,d}(x) = \frac{1}{\lambda_{\text{speech}} + s\mu_{\text{speech}} + \lambda_{\text{video}} + v\mu_{\text{video}} + \lambda_{\text{data}} + \beta_{\text{data}}\left(s,v,d\right)\left(d_t\left(s,v,d\right)-1\right)\mu_{\text{data}}} \cdot \left\{\begin{array}{c} 1 + \lambda_{\text{speech}}\widehat{\tau}_{s+1,v,d}(x) + s\mu_{\text{speech}}\widehat{\tau}_{s-1,v,d}(x) + \lambda_{\text{video}}\widehat{\tau}_{s,v+1,d}(x) + v\mu_{\text{video}} \cdot \\ \widehat{\tau}_{s,v-1,d}(x) + \lambda_{\text{data}}\widehat{\tau}_{s,v,d+1}(x) + \beta_{\text{data}}\left(s,v,d\right)\left(d_t\left(s,v,d\right)-1\right)\mu_{\text{data}}\widehat{\tau}_{s,v,d-1}(x) \end{array}\right\}.$$

Let the auxiliary matrices $\widehat{\mathcal{Q}}^\star_{00}$ $(\widetilde{\mathcal{Q}}^\star_{00})$ contain the (off-)diagonal entries of $\mathcal{Q}^\star_{00}$ and zeros elsewhere so that $\mathcal{Q}^\star_{00} = \widehat{\mathcal{Q}}^\star_{00} + \widetilde{\mathcal{Q}}^\star_{00}$. Then the above set of expressions can be written in matrix form as

$$\widehat{\boldsymbol{\tau}}_0(x) = \left(-\widehat{\mathcal{Q}}^\star_{00}\right)^{-1} \left\{\mathbf{1} + \widetilde{\mathcal{Q}}^\star_{00}\widehat{\boldsymbol{\tau}}_0(x) + \mathcal{Q}^\star_{0+}\widehat{\boldsymbol{\tau}}_+(x)\right\}$$

$$\Leftrightarrow \widehat{\boldsymbol{\tau}}_0(x) = \left(-\mathcal{Q}^\star_{00}\right)^{-1} \left\{\mathbf{1} + \mathcal{Q}^\star_{0+}\widehat{\boldsymbol{\tau}}_+(x)\right\}, \qquad (3.14)$$

where the vector $\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathbf{1}$ contains the expected accumulated delay until the system moves to a state $(s,v,d) \in \mathbb{S}^+_{\text{data},+}$, and probability matrix $\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}$ gives for each state $(s,v,d) \in \mathbb{S}^+_{\text{data},0}$ (no data transfer) the probability distribution over the states in $\mathbb{S}^+_{\text{data},+}$ where the system enters a state of actual data transfer again. Note that $\mathcal{Q}^\star_{00}$ is the infinitesimal generator of a transient Markov chain, so that it is non-singular and thus invertible.

Next we concentrate on the evolution from state $(s,v,d) \in \mathbb{S}^+_{\text{data},+}$. Consider a time interval of length $\Delta > 0$, with Δ sufficiently small such that the tagged data call cannot terminate within this time, hence $\Delta < x/\beta_{\text{data}}(s,v,d)$ in state $(s,v,d) \in \mathbb{S}^+_{\text{data},+}$ (recall that $\beta_{\text{data}}(s,v,d) > 0$ for all $(s,v,d) \in \mathbb{S}^+_{\text{data},+}$). Condition on all the possible events occurring in this interval, starting in state $(s,v,d) \in \mathbb{S}^+_{\text{data},+}$. For notational convenience and readability, the boundary constraints are not explicitly considered. Equations for the boundary can be derived analogously.

$$\begin{aligned}
&\widehat{\tau}_{s,v,d}(x)\\
&\quad = \quad \Delta\\
&\qquad +\lambda_{\text{speech}}\Delta\widehat{\tau}_{s+1,v,d}(x - O(\Delta))\\
&\qquad +s\mu_{\text{speech}}\Delta\widehat{\tau}_{s-1,v,d}(x - O(\Delta))\\
&\qquad +\lambda_{\text{video}}\Delta\widehat{\tau}_{s,v+1,d}(x - O(\Delta))\\
&\qquad +v\mu_{\text{video}}\Delta\widehat{\tau}_{s,v-1,d}(x - O(\Delta))\\
&\qquad +\lambda_{\text{data}}\Delta\widehat{\tau}_{s,v,d+1}(x - O(\Delta))\\
&\qquad +\beta_{\text{data}}(s,v,d)\left(d_t(s,v,d) - 1\right)\mu_{\text{data}}\Delta\widehat{\tau}_{s,v,d-1}(x - O(\Delta))\\
&\qquad +(1 - (\lambda_{\text{speech}} + s\mu_{\text{speech}} + \lambda_{\text{video}} + v\mu_{\text{video}} + \lambda_{\text{data}} + \beta_{\text{data}}(s,v,d)\\
&\qquad\qquad \times \left(d_t(s,v,d) - 1\right)\mu_{\text{data}})\Delta)\widehat{\tau}_{s,v,d}(x - \beta_{\text{data}}(s,v,d)\Delta)\\
&\qquad +o(\Delta).
\end{aligned}$$

Rearranging terms and letting $\Delta \downarrow 0$ gives the system of differential equations

$$\begin{aligned}
&\beta_{\text{data}}(s,v,d)\,\frac{\partial}{\partial x}\widehat{\tau}_{s,v,d}(x)\\
&\quad= \quad 1+\lambda_{\text{speech}}\widehat{\tau}_{s+1,v,d}(x)+s\mu_{\text{speech}}\,\widehat{\tau}_{s-1,v,d}(x)+\lambda_{\text{video}}\widehat{\tau}_{s,v+1,d}(x)\\
&\qquad\quad +v\mu_{\text{video}}\widehat{\tau}_{s,v-1,d}(x)+\lambda_{\text{data}}\widehat{\tau}_{s,v,d+1}(x)+\beta_{\text{data}}(s,v,d)\,(d_t(s,v,d)-1)\\
&\qquad\quad \times\mu_{\text{data}}\widehat{\tau}_{s,v,d-1}(x)+(-\lambda_{\text{speech}}-s\mu_{\text{speech}}-\lambda_{\text{video}}-v\mu_{\text{video}}-\lambda_{\text{data}}\\
&\qquad\quad -\beta_{\text{data}}(s,v,d)\,(d_t(s,v,d)-1)\,\mu_{\text{data}})\widehat{\tau}_{s,v,d}(x).
\end{aligned}$$

Note that since $\beta_{\text{data}}(s,v,d)>0$ for all $(s,v,d)\in\mathbb{S}^{+}_{\text{data},+}$ the limit taken for $\Delta\downarrow 0$ is well-defined. For the considered partition of the state space, the above system of differential equations for the vector of conditional expected transfer times $\widehat{\boldsymbol{\tau}}_{+}(x)$, can be written in matrix form as follows:

$$\mathcal{B}_{+}\,\frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}_{+}(x)=\mathbf{1}+\mathcal{Q}^{\star}_{+0}\widehat{\boldsymbol{\tau}}_{0}(x)+\mathcal{Q}^{\star}_{++}\widehat{\boldsymbol{\tau}}_{+}(x), \tag{3.15}$$

with initial condition

$$\widehat{\boldsymbol{\tau}}_{+}(0)=\mathbf{0}. \tag{3.16}$$

The initial condition (3.16) reflects that the transfer time $\tau_{s,v,d}(0)$ of an 'empty' data call starting in system state $(s,v,d)\in\mathbb{S}^{+}_{\text{data},+}$ is zero, almost surely. Substitution of (3.14) replaces (3.15) by

$$\begin{aligned}
\frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}_{+}(x)&=\mathcal{B}_{+}^{-1}\left\{\mathbf{1}+\mathcal{Q}^{\star}_{+0}\left[\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\left(\mathbf{1}+\mathcal{Q}^{\star}_{0+}\widehat{\boldsymbol{\tau}}_{+}(x)\right)\right]+\mathcal{Q}^{\star}_{++}\widehat{\boldsymbol{\tau}}_{+}(x)\right\}\\
=\mathcal{B}_{+}^{-1}&\left\{\mathcal{I}+\mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\right\}\mathbf{1}+\mathcal{B}_{+}^{-1}\left\{\mathcal{Q}^{\star}_{++}+\mathcal{Q}^{\star}_{+0}\left(-\mathcal{Q}^{\star}_{00}\right)^{-1}\mathcal{Q}^{\star}_{0+}\right\}\widehat{\boldsymbol{\tau}}_{+}(x). \qquad (3.17)
\end{aligned}$$

The remainder of the proof is similar to that of Proposition 3.1. The existence and uniqueness of a solution $\widehat{\boldsymbol{\tau}}_{+}(x)$ to a system of differential equations of the form (3.17) for every initial vector, is a known result, see e.g. [51, Chapter 1, Section 8]. The existence and uniqueness of a vector $\boldsymbol{\gamma}_{\text{data}}$ that satisfies (3.10), (3.11) is guaranteed by standard Markov reward chain theory, see e.g. [213, Theorem 3.1, page 167]. The proposition is then proven by substituting the claimed unique solution (3.12) into the system of differential equations and verifying the initial condition.

In order to prove the asymptotic result (3.13), note that since $\mathcal{B}_+^{-1}\left\{\mathcal{Q}^\star_{++} + \mathcal{Q}^\star_{+0}\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}\right\}$ is the infinitesimal generator of an irreducible finite state space Markov chain with equilibrium distribution vector $\frac{\boldsymbol{\pi}^\star_{\text{data},+}\mathcal{B}_+}{\boldsymbol{\pi}^\star_{\text{data},+}\mathcal{B}_+\mathbf{1}}$, it holds that

$$\lim_{x\to\infty}\left[\mathcal{I} - \exp\left\{x\mathcal{B}_+^{-1}\left(\mathcal{Q}^\star_{++} + \mathcal{Q}^\star_{+0}\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}\right)\right\}\right]\boldsymbol{\gamma}_{\text{data}} = \boldsymbol{\gamma}_{\text{data}},$$

using (3.11). The asymptotic derivation for $\widehat{\boldsymbol{\tau}}_0(x)$ further uses the fact that $\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}$ is a probability matrix, so that

$$\left(-\mathcal{Q}^\star_{00}\right)^{-1}\mathcal{Q}^\star_{0+}\left(\frac{x}{\boldsymbol{\pi}^\star_{\text{data},+}\mathcal{B}_+\mathbf{1}}\mathbf{1}\right) = \frac{x}{\boldsymbol{\pi}^\star_{\text{data},+}\mathcal{B}_+\mathbf{1}}\mathbf{1}.$$

□

Remark 3.3 A comparison of the marginal analysis used in the proof of Proposition 3.1 to derive the conditional QOS of video calls, and that used in the proof of Proposition 3.4 above to derive the conditional QOS of data calls, reveals noteworthy similarities. Whereas in both cases we start from a given system state and look a small period Δ ahead in time, in the data call QOS analysis the cumulative *transfer time* $\widehat{\tau}_{s,v,d}(x)$ directly increases by Δ, while in the video call QOS analysis the increase of the cumulative *transfer volume* $\widehat{x}_{s,v,d}(\tau)$ is influenced by the starting state and the events that may happen during Δ. The manner in which the channel assignment functions $\beta_{\text{video}}(s,v,d)$ and $\beta_{\text{data}}(s,v,d)$ appear in QOS expressions (3.7) and (3.12), respectively, reflect the intuitively obvious fact that a more generous channel assignment increases the expected video throughput ($\sim \widehat{\mathbf{x}}(\tau)$), or decreases the expected data call transfer time ($\sim \widehat{\boldsymbol{\tau}}(x)$).

Remark 3.4 The asymptotic linearity of $\widehat{\boldsymbol{\tau}}(x)$ and hence of $\mathbf{T}_{\text{data}}(x)$ in x indicates a fairness property. Aside from a basic sojourn time component corresponding to a.o. the size-independent access time, the size-dependent component is approximately a constant factor times the call size x.

3.6.2. SPEECH/HIGH-/LOW-PRIORITY DATA MODEL

Also for the SHL model, we can derive the expected sojourn times of high- and low-priority data calls, conditional on the call size and the system state upon arrival. As the conditional analysis is similar to that given for data calls in the SVD model considered above, the lengthy derivations are omitted here for reasons of concision. We merely summarise the set of conditional performance measures that can be obtained:

HIGH-PRIORITY DATA SERVICE		LOW-PRIORITY DATA SERVICE	
$\mathbf{T}_{\text{high}}(x)$	for $x \geq 0$	$\mathbf{T}_{\text{low}}(x)$	for $x \geq 0$
$\widehat{\sigma}_{\text{high},s,h,\ell}(x)$	for $x \geq 0$, $(s,h,\ell) \in \mathbb{S}^{+}_{\text{high}}$	$\widehat{\sigma}_{\text{low},s,h,\ell}(x)$	for $x \geq 0$, $(s,h,\ell) \in \mathbb{S}^{+}_{\text{low}}$
$\widehat{\tau}_{\text{high},s,h,\ell}(x)$	for $x \geq 0$, $(s,h,\ell) \in \mathbb{S}^{+}_{\text{high}}$	$\widehat{\tau}_{\text{low},s,h,\ell}(x)$	for $x \geq 0$, $(s,h,\ell) \in \mathbb{S}^{+}_{\text{low}}$
Ψ_{high}		Ψ_{low}	

3.7. NUMERICAL RESULTS

This section presents an extensive numerical study in order to demonstrate the relevant trade-offs between the GOS and QOS of the different services. As it would take too much space to study the effect of all model parameters, some parameters are prefixed at a realistic level while the remaining parameters are varied within a realistic range around their default value and their impact on the relevant performance measures is investigated. Table 3.2 below gives an overview of all system and traffic model parameters and indicates either their prefixed value or a range of considered values. Some comments regarding these parameters are made below.

Regarding the *system parameters*, the GSM/GPRS cell capacity is prefixed by a typical assignment of 3 frequencies, corresponding with $3 \times 8 = 24$ physical channels, of which $C_{\text{total}} = 22$ are traffic channels, while the remaining 2 are used for control signalling. In the corresponding models, a range of C_{speech}, C_{video} and C_{data} is considered to assist a network operator in deploying appropriate channel sharing policies. A GPRS terminal is expected to have multichannel capability $\beta^{\max}_{\text{GPRS}} = 4$. In each of the numerical experiments, the access queue sizes (Q_a, $Q_{a,\text{high}}$, $Q_{a,\text{low}}$) have been set sufficiently large to ensure a negligible amount of data call blocking, while the transfer queue is dimensioned at $Q_t \equiv \lceil C_{\text{total}}/\beta^{\max}_{\text{GPRS}} \rceil = 6$ data calls (of each priority class), which allows just enough active data calls to occupy the number of traffic channels that are potentially available for data transfer and thus establish a work-conserving

Table 3.2 Numerical results: parameter settings.

PARAMETER	PREFIXED VALUE		PARAMETER	DEFAULT VALUE		RANGE
C_{total}	22	channels	C_{speech}	14	channels	$\{6, 10, 14\}$
$\beta_{\text{GPRS}}^{\text{max}}$	4	channels	C_{video}	6	channels	$\{2, 4, \cdots, 16\}$
Q_a	∞	calls	C_{data}	2	channels	$\{0, 2, \cdots, 14\}$
$Q_{a,\text{high}}$	∞	calls	ϕ	0		$[0, 1]$
$Q_{a,\text{low}}$	∞	calls	ρ_{video}	0.4555	Erlang	$(0, 20]$
Q_t	6	calls	ρ_{data}	2.2000	Erlang	$\{2.2, 4.4, 6.6\}$
μ_{speech}	0.0200	calls/s				
ρ_{speech}	13.6513	Erlang				
μ_{video}	0.0200	calls/s				
$\beta_{\text{video}}^{\text{min}}$	2	channels				
r_{video}	13.4	kbits/s				
μ_{data}	0.0283	calls/s				
r_{data}	9.05	kbits/s				
ξ	50%					

property. In the SHL model, the default value of the DPS service weight ϕ is 0, turning the low-priority class into a best effort class.

With regard to the service parameters, the average speech and video call durations are both equal to 50 seconds. The speech traffic load ρ_{speech} Erlang is chosen such that for a cell with 3 frequencies the speech call blocking probability is 1% provided that all C_{total} channels are available for speech service. The default video traffic load ρ_{video} is set to 0.4555 Erlang, corresponding with a 1% video call blocking probability given a capacity of $C_{\text{video}} = 6$ channels and a minimum video call assignment of $\beta_{\text{video}}^{\text{min}} = 2$ channels. The assumed information bit rate $r_{\text{video}} = 13.4$ kbits/s of each video traffic channel is based on GPRS channel coding scheme CS-2, as it offers a good balance between channel throughput and bit protection. The data call size parameter $\mu_{\text{data}} = (320/9.05)^{-1}$ seconds corresponds with an average e-mail size of 320 kbits that is transferred over a single ($r_{\text{data}} =$) 9.05 kbits/s traffic channel (CS-1 is assumed for its maximum error correction capability). The default data traffic load is set to 2.2 Erlang, or 10% of the cell's aggregate transfer capacity of 22 Erlang, while $\rho_{\text{data}} \in \{4.4, 6.6\}$ Erlang are used as alternatives. In the SHL model, $\xi = 50\%$ of all data calls have high priority.

In the remainder of this section, a number of numerical experiments is presented in order to obtain insight in the effect of the variable parameters on the performance measures. Regarding the applied units, it is noted that the data call access, transfer and sojourn times are expressed in *seconds*, the video throughputs are expressed in *kbits/s*, the data call size is expressed in units of r_{data} kbits, and the video call size (duration) is expressed in *seconds*. The experiments are organised per system model, starting with the SVD model.

3.7.1. SPEECH/VIDEO/DATA MODEL

The channel sharing policy proposed for the SVD model is evaluated assuming default settings for all traffic model parameters while varying the system model parameters C_{speech}, C_{video} and C_{data}. The channel utilisation $\mathbf{U}$ and GOS performance measures $\mathbf{P}_{\mathrm{speech}}$ and $\mathbf{P}_{\mathrm{video}}$ are depicted in Figure 3.5. Recall that the data call access queue Q_a was dimensioned sufficiently large to ensure that $\mathbf{P}_{\mathrm{data}} \approx 0$. As the figure shows, the choice of C_{data} greatly influences both the channel utilisation and call blocking. Although these reserved channels are available for video transfer unless needed for data transfer, they provide no basis for video or speech call admission. On the other hand, the split of the remaining $C_{\mathrm{total}} - C_{\mathrm{data}}$ channels into subsets of C_{speech} and C_{video} channels, has relatively little impact on $\mathbf{U}$, $\mathbf{P}_{\mathrm{speech}}$ or $\mathbf{P}_{\mathrm{video}}$, except for the rather extreme case where $C_{\mathrm{speech}} = 14$, $C_{\mathrm{data}} = 6$ and thus $C_{\mathrm{video}} = 2$, where only a single video call can be admitted and thus a great many are blocked.

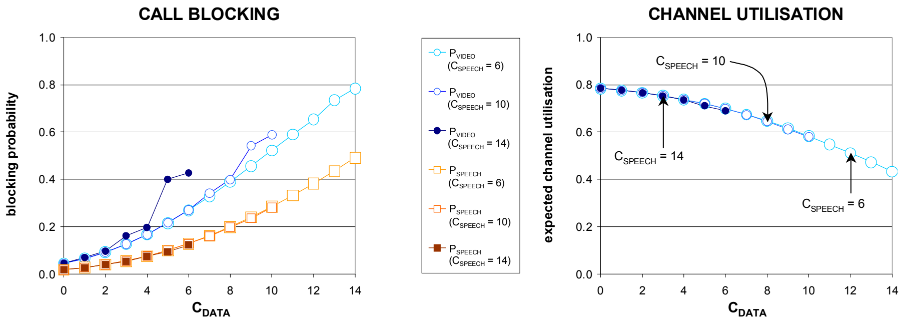


Figure 3.5 GOS of speech and video calls (left) and expected channel utilisation (right) versus C_{data} in the SVD model for $C_{\mathrm{speech}} \in \{6, 10, 14\}$.

The QOS experienced by video and data calls is illustrated by the numerical results displayed in Figure 3.6. The QOS of data calls converges to the absolute minimum expected sojourn time of 320 kbits / ($r_{\text{data}}\,\beta_{\text{GPRS}}^{\text{max}}$ kbits/s) $\approx$ 8.84 seconds as C_{data} increases, while it is relatively independent of C_{speech} or C_{video}. The marginal data QOS improvement of a reservation greater than $C_{\text{data}} = 4$ channels seems hardly worth the induced cost primarily in terms of the speech and video call blocking probabilities.

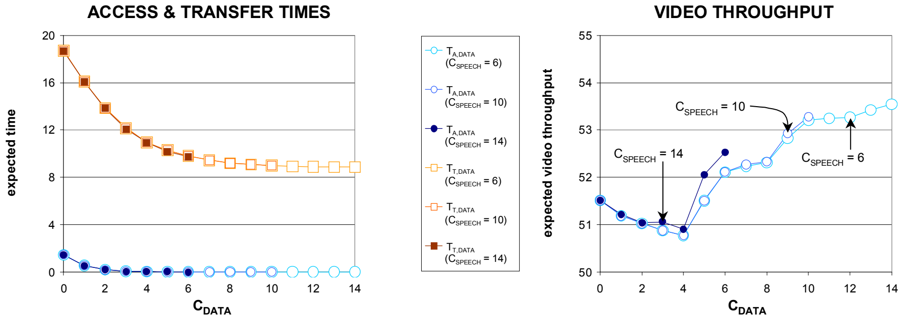


Figure 3.6 Expected QOS of data (left) and video (right) calls versus C_{data} in the SVD model for $C_{\text{speech}} \in \{6, 10, 14\}$.

The time-average video throughput $\mathbf{R}_{\text{video}}^{t}$ (recall that this is defined *conditional* on the presence of at least one video call) varies within a narrow interval lower bounded by 50.8 kbits/s and upper bounded by the theoretical maximum of $r_{\text{video}}\,\beta_{\text{GPRS}}^{\text{max}} = 53.6$ kbits/s, and is characterised by a rather irregular dependency on C_{data}. Consider for instance the curve corresponding with $C_{\text{speech}} = 6$. As C_{data} is raised from 0 to 14, C_{video} decreases accordingly from 16 to 2, which induces two opposite effects: *(i)* a smaller base capacity exists for video (and speech) call admission so that $\mathbf{P}_{\text{video}}$ increases (see Figure 3.5) and the degree of competition for resources among *admitted* video calls is reduced, leading to an improvement of the video QOS; *(ii)* at the same time, the exchange of preferred video for reserved data channels implies that the idle capacity on these channels, originally available to upgrade video call quality, is now assigned to data calls, thus deteriorating the video QOS. Not surprisingly, for the given load parameters, the latter effect outweighs the former effect for $C_{\text{data}} \leq 4$, i.e. precisely where the data calls show the biggest room for QOS improvement. For $C_{\text{data}} > 4$, the expected data QOS is so close to the theoretical optimum, that the exchanged channels remain available to the lower number of admitted video calls, so that the former effect dominates.

Regarding the shape of the video QOS curve we further note that as C_{data} increases, video calls have to grab their non-guaranteed traffic channels more and more among the C_{data} reserved data traffic channels (rather then among the $C_{\text{speech}} + C_{\text{video}}$ traffic channels), which implies that the *instantaneous* video QOS is likely to vary more burstily over time, since data call arrivals and departures generate a batch-wise variation in the availability of traffic channels, as opposed to speech calls in the $C_{\text{speech}} + C_{\text{video}}$ channel regime. In combination with the *on average* higher channel availability per admitted video call due to the lower number of admitted speech and video calls, this leads to a the slightly irregular relation between the expected video QOS and C_{data}. In passing, we note that the probability that $\beta_{\text{GPRS}}^{\max}$ limits the instantaneous video QOS exhibits a very similar dependency on C_{data} (not shown).

For the channel partitioning of $(C_{\text{speech}}, C_{\text{video}}, C_{\text{data}}) = (14, 6, 2)$, Figure 3.7 demonstrates the conditional QOS of both video and data calls as a function of the respective call sizes, as well as the corresponding asymptotes that have been derived. In the included legend, '$\mathbf{T}_{\text{data}}, \mathbf{R}^{c}_{\text{video}}$' indicates that the associated marker refers to the expected data call sojourn time in the left chart and to the call-average video throughput in the right chart, a 'legend sharing' method that will be applied in later figures as well. As $C_{\text{data}} > 0$ and thus $\mathbb{S}^{+}_{\text{data},0} = \emptyset$, a simpler form of Proposition 3.4 has been applied to obtain the conditional expected data call transfer times. The displayed video call duration ranges from 0 to 250 seconds, while the data call size is recalled to be expressed in units of r_{data} kbits and thus ranges from 0 to $250\, r_{\text{data}} \approx 2263$ kbits. The upper bound of the call size range is for both services approximately equal to the 99$^{\text{th}}$ percentile of the corresponding call size distribution. Note that for both services, the exact QOS approaches the derived asymptotes already for relatively small video call durations and data call sizes, which demonstrates the value of the results of Section 3.6.

In addition to $\mathbf{R}^{c}_{\text{video}}(\tau)$, $\mathbf{T}_{\text{data}}(x)$ and their respective asymptotes, the figure also shows the best-case (video: $\widehat{x}_{0,1,0}(\tau)/\tau$ and data: $\widehat{\sigma}_{0,0,1}(x)$) and worst-case curves (video only: $\widehat{x}_{14,3,\infty}(\tau)/\tau$) of the conditional expected QOS, given an empty or full system upon arrival of the considered call, respectively (see dashed curves). The worst-case conditional expected video throughput $\widehat{x}_{14,3,\infty}(\tau)/\tau$ is readily calculated using the SV(D) model without data traffic and setting $C_{\text{total}} = 20$ rather than 22, since the infinite number of data calls ensures continuous use of the 2 reserved data

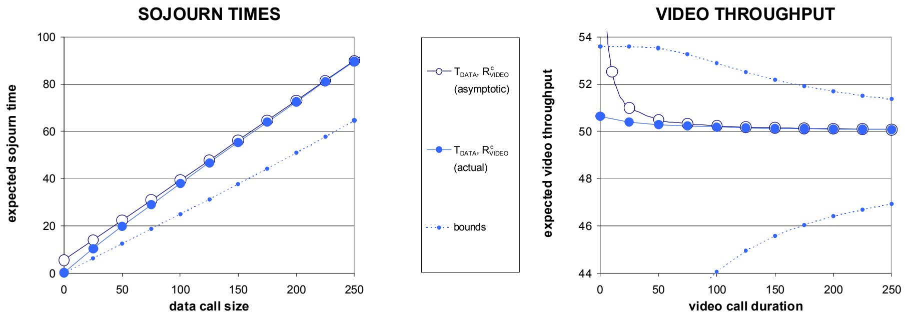


Figure 3.7 Conditional expected QOS of data (left) and video (right) calls versus the elastic call size in the SVD model.

traffic channels. In contrast, since $Q_a = \infty$, the worst-case conditional expected data call sojourn times $\widehat{\sigma}_{s,v,\infty}(x) = \infty$, regardless of s and v, and are thus not included in the figure. Note that for the video service, the best- and worst-case curves converge towards the average curve, since the impact of the initial state fades out. In contrast, this is not the case for the data service, since the significant throughput gain that can be achieved initially when a data call finds the system empty upon arrival, directly affects its service requirement (remaining call size) and shortens its sojourn time.

As a final experiment of this subsection, consider the specific incidence of the SVD model with $C_{\text{data}} = 0$, in which case both the speech GOS and the video GOS and QOS are unaffected by the data traffic. An intermediate result in the determination of the conditional expected (call-average) video throughput values $\mathbf{R}^{c}_{\text{video}}(\tau)$ is given by the values of $r_{\text{video}}\widehat{x}_{s,v,d}(\tau)/\tau$, the conditional expected throughput of an admitted video call of duration τ arriving in system state $(s, v, d) \in \mathbb{S}^{+}_{\text{video}}$ (recall that v includes the new call). This result may be very useful as a feedback information service to the caller. Figure 3.8 presents an illustrative example of $r_{\text{video}}\,\widehat{x}_{s,v,d}(\tau)/\tau$ versus $(s, v, \cdot) \in \mathbb{S}^{+}_{\text{video}}$ (independent of d) for an admitted video call of duration 50 seconds, given $C_{\text{speech}} = 14$, $C_{\text{video}} = 8$, and all other parameters at their default setting. Note the domain $\left\{(s, v, \cdot) : s + \beta^{\min}_{\text{video}}\, v \leq C_{\text{total}} \text{ and } v \leq C_{\text{video}}/\beta^{\min}_{\text{video}} = 4\right\}$. The figure supports the intuition that $r_{\text{video}}\,\widehat{x}_{s,v,d}(\tau)/\tau$ is decreasing in both s and v, i.e. that the conditional expected video throughput is lower as the video call finds the cell to be more congested upon its arrival. The numerical example further illustrates that the expected video throughput is most sensitive to a change in the number of *video* calls, since an additional video call claims $\beta_{\text{video}}^{\min} = 2$ times as many traffic channels as an additional speech call.

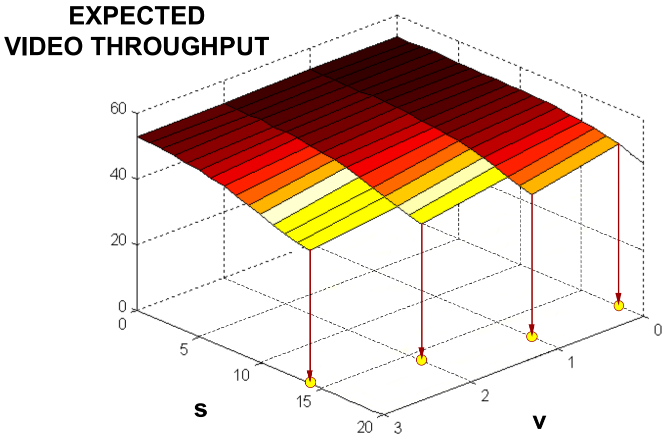


Figure 3.8 Conditional expected QOS of video calls versus the system state upon call admission in the SVD model.

3.7.2. SPEECH/HIGH-/LOW-PRIORITY DATA MODEL

In the SHL model serving speech and two priority classes of data calls, the principal instrument for data QOS differentiation is the service weight factor $\phi \in [0, 1]$ of the DPS scheduling scheme, defined as the relative amount of attention given to a low-priority data call compared to a high-priority data call. Recall that under $\phi = 0$ the low-priority class becomes a best effort class, while at the other extreme $\phi = 1$ corresponds with *no* QOS differentiation, i.e. equal treatment of both priority classes. For $C_{\text{data}} = 0$, three different data traffic loads, and otherwise default parameter settings (see Table 3.2), Figure 3.9 depicts the expected sojourn time of high- and low priority data calls as a function of ϕ.

We note that the expected channel utilisation and the call blocking probabilities (not displayed) are independent of the choice of ϕ, where it is recalled that the access queue sizes have been set sufficiently large to ensure a negligible amount of data call blocking. The numerical results illustrate that the discrepancy in QOS between high- and low-priority data calls is maximal for $\phi = 0$, while it diminishes as ϕ grows until the QOS levels are identical for $\phi = 1$. Furthermore, as higher data traffic loads induce a more intense competition for the varying spare capacity, the difference in QOS levels is also more significant. Suppose a network operator (or service provider) deems QOS differentiation (and correspondingly: differentiated charging) only feasible if it can establish a QOS difference of at least a factor 2, then the presented results indicate that only for $\rho_{\text{data}} = 4.4$ and $\rho_{\text{data}} = 6.6$ this requirement can be met, choosing $\phi \approx 0$ and $\phi \lessapprox 0.5$, respectively.

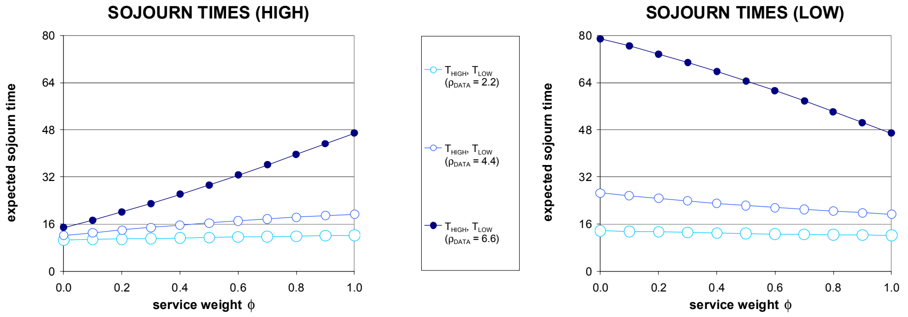


Figure 3.9 Expected QOS of high- (left) and low-priority (right) data calls versus the scheduling weight ϕ in the SHL model for $\rho_{\text{data}} \in \{2.2, 4.4, 6.6\}$.

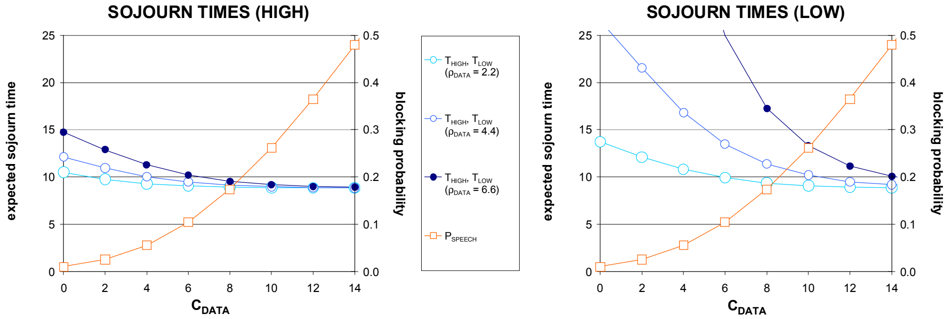


Figure 3.10 Expected QOS of high- (left) and low-priority (right) data calls versus C_{data} in the SHL model for $\rho_{\text{data}} \in \{2.2, 4.4, 6.6\}$.

Although the dedication of cell capacity to the data service by means of C_{data} is not so much a means to affect the degree of QOS differentiation, but rather to establish an overall QOS improvement, the results in Figure 3.10 illustrate that the choice of C_{data} does influence the relative as well as the absolute QOS levels (assuming $\phi = 0$). A higher reservation level augments the amount of available resources and hence reduces the expected sojourn times for both priority classes. All curves tend to converge towards the lower bound of 320 kbits / $(r_{\text{data}}\,\beta^{\text{max}}_{\text{GPRS}}) \approx 8.84$ seconds. In order to optimise C_{data} an operator must balance the corresponding speech call blocking probability $\mathbf{P}_{\text{speech}}$ (see also Figure 3.10) and the data QOS according to its service policy. Subsequently, the service weight parameter ϕ is to be set in order to establish the desired level of QOS differentiation (see above).

3.7.3. SPEECH/DATA MODEL: TRANSIENT CORRELATION

The final stand-alone numerical experiment presented in this chapter is a simulation experiment taken from [126], which is incorporated to illustrate the *transient* correlation between the data call sojourn time and the number of present speech calls. Speech calls and a single class of data calls, which model individual IP packets with a fixed size of 536 bytes, are integrated in a single GSM/GPRS cell with a capacity of $C_{\text{total}} = 7$ traffic channels (one frequency). The offered speech traffic load equals $\rho_{\text{speech}} = 1.88$ Erlang. Figure 3.11 shows how the data call sojourn time and the number of speech calls evolve over time in a typical simulation trace when the ρ_{data} (left chart) and C_{data} (right chart) are varied. The underlying traces of generated speech calls are identical in all cases, while the data call trace underlying the right chart is identical to that corresponding with the curve for $\rho_{\text{data}} = 2.8$ Erlang in the left chart. We note that although the system and traffic parameters differ from earlier numerical experiments, the demonstrated qualitative trends are similar.

As the charts reveal, a clear correlation exists between the data call sojourn time and the number of speech calls, which can be observed from the fact that the peaks of the data call sojourn time coincide with or immediately follow the peaks of the number of present speech calls. Observe that the correlation weakens as the data calls' need to use the shared channels is smaller, e.g. lower ρ_{data} (left chart: compare the cases of 0.7 and 2.8 Erlang data traffic load) or more dedicated channels (right chart: compare $C_{\text{data}} = 0$ with $C_{\text{data}} = 2$), since the capacity used by data calls becomes less dependent on the number of (prioritised) speech calls present. Observe that the expected data QOS improvement under a lower ρ_{data} or a higher C_{data}, is apparent only at the occasional instances of congestion, while at other times, the

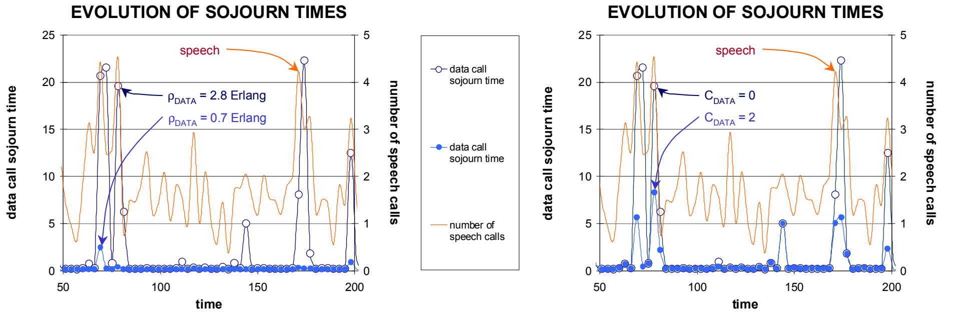


Figure 3.11 The transient correlation between the data call sojourn time and the number of speech calls present for different ρ_{data} (left, for $C_{\text{data}} = 0$) and C_{data} (right, for $\rho_{\text{data}} = 2.8$ Erlang).

discrepancy between the curves is negligible and the data call sojourn time is limited by the terminals' multislot capability only. Note that the short peak in the data QOS curve around 140 seconds is caused by an incidentally large number of competing data calls in the system, while the considered channel reservation imposed by $C_{\text{data}} = 2$ has no effect due to the relatively small number of speech calls present.

3.8. CONCLUDING REMARKS

We have developed and extensively analysed a generic model for performance evaluation, parameter optimisation and dimensioning of an integrated GSM/GPRS network serving speech, video and data calls, potentially offered in distinct priority classes. While speech calls are characterised by a fixed single channel assignment, the scalable video calls and delay-tolerant data calls are elastic in the sense that they can handle a channel assignment that varies between an optional minimum guaranteed rate and the upper bound imposed by the GPRS terminals' multislot capability. The principal difference between both types of elastic calls lies in the impact of the varying channel assignment on the experienced QOS. For video calls, which are characterised by an autonomously exponentially distributed duration, the experienced channel assignments influence the average throughput and hence the perceived audio/video quality. Data calls comprise of a file transfer with an exponentially distributed size, whose sojourn time depends on the (varying) channel assignment. While the presented generic model and analysis in principle allow treatment of a broader variety of scenarios, we have targeted the performance analysis and numerical section at two distinct cases for reasons of enhanced transparency: the SVD model integrating speech, video and data calls, and the SHL model, integrating speech calls and two priority classes of data calls, where a Discriminatory Processor Sharing scheduling discipline is deployed to establish data QOS differentiation.

Markov chain analysis has been applied to derive basic performance measures such as the expected channel utilisation, service-specific blocking probabilities, and QOS measures for the elastic services, i.e. expected video throughput and expected (priority class-specific) data call sojourn times. Furthermore, exact expressions were derived for the expected video and data QOS, conditional on the call duration or file size, respectively, and the system state of arrival. Aside from providing insight in the fairness that is provided among calls of a given service type but with different durations or sizes, as a potential application these conditional performance measures may be fed back to the caller at call establishment as an indication of the expected QOS.

An extensive numerical section has been included to illustrate the effects of GOS and QOS differentiation schemes and traffic load parameters on system performance, providing a basis for optimisation of operator-controlled system parameters. For instance, we have provided insight in the impact of channel reservation policies on the performance of the different services, shown the impact of the prioritisation weight in the SHL model and demonstrated the attainability of conditional performance measures.

With regard to data services, both the previous and the present chapter have concentrated on the sojourn time analysis of data calls with exponentially distributed sizes. Chapters 4 and 5 present supplemental investigations that focus on the impact of the data *call size variability* and an investigation of *throughput measures*, respectively, where the latter chapter is not only targeted at data services but also includes a more extensive video throughput analysis.

CHAPTER 4

SENSITIVITY ANALYSIS OF AN INTEGRATED SERVICES MODEL

ANALYTICAL performance evaluation studies of integrated services networks typically assume Markovian models for reasons of mathematical tractability (e.g. [14, 175], Chapters 2 and 3 of this monograph). In particular, data call sizes are assumed to be exponentially distributed. Whereas the exponential distribution is rather light-tailed with a coefficient of variation equal to one, it is commonly acknowledged that e.g. WWW pages are *heavy-tailed* [56, 58, 143]. The objective of this chapter is to investigate the sensitivity of the data performance with respect to the data call size variability in an integrated services GSM/GPRS network. It is stressed however, that the considered models are sufficiently generic to allow broader application of the principal conclusions.

We consider a single GSM/GPRS cell integrating a circuit-oriented speech telephony service, requesting a fixed capacity assignment, and a delay-tolerant packet-oriented elastic data service (e.g. file transfer). As the speech calls are assumed to have strict and preemptive priority over data calls, a varying amount of capacity is available to serve the data calls. At any time, the available capacity is fairly shared by the present data calls according to a Processor Sharing service discipline. The principal performance measure of interest is the Quality of Service of data calls, expressed in terms of the *expected sojourn time*.

A series of dynamic simulations of the considered model is presented in order to investigate the impact of the data call size distribution on the experienced QOS, demonstrating that *the greater the data call size variability, the better the* QOS. An intuitive explanation for this remarkable phenomenon can be formulated as follows. In a processor sharing model with varying capacity, tiny data calls are argued to suffer relatively little from the capacity fluctuations, as they are generally swiftly transferred, experiencing never more than perhaps a single period without any service. At the other extreme, a very large data call effectively 'sees' a system with fixed capacity, as the capacity fluctuations tend to average out within its lifetime. The medium-sized calls in between these extremes are the ones that are likely to suffer most from the capacity fluctuations. With an increasing variability of the data call size distribution, while keeping its mean fixed, the relative number of tiny data calls increases, while the large data calls become larger and thus less sensitive to the capacity fluctuations. Since the considered QOS measure is a *call*-average measure, the degraded QOS of the medium-sized data calls is more and more outweighed by the relatively favourable QOS of the growing multitude of small calls.

In order to illustrate the significance of the presented phenomenon, we note that in a (fixed capacity) FIFO queue the effect is known to be reversed: the QOS *degrades* under a greater call size variability (depending only on its first and second moments (Pollaczek-Khintchine formula, see e.g. [213]), as a rare large data call in service causes huge queueing delays. A similar trend is argued and demonstrated to hold for a FIFO queue with varying capacity (see Section 4.6). In contrast, the data call sojourn times in a fixed capacity PS queue are known to be *insensitive* to the call size variability [203, 204, 213]. The contribution of this chapter is to demonstrate and analytically support the phenomenon that in the practically most relevant PS queue with varying capacity (and an unlimited number of service positions), the QOS *improves* under a greater call size variability. The presented numerical results indicate however that it is not trivial which data call size variability measure captures the essence of its impact on the QOS. Within a family of probability distributions, both the second moment of the data call size and the heaviness of the distribution tail are observed to suffice, while these variability measures turn out to be less useful in a comparison across distinct families of distributions.

The outline of this chapter is as follows. Section 4.1 reviews the relevant literature, directly followed by a statement of contribution in Section 4.2. The considered integrated services model is described in Section 4.3, featuring a PS service discipline for admitted data calls. Section 4.4 presents the observation that the expected data call sojourn time is decreasing in the variability of the data call size distribution. Subsequently, Section 4.5 provides theoretical support and intuition based on an analysis of extreme cases. As is demonstrated in Section 4.6, the QOS improvement with greater call size variability is specifically due to the PS nature of the service process. In particular, in Section 4.6 we explicitly investigate the trade-off between the FIFO and PS service disciplines and study the impact of the data call size variability on the QOS in an extended model that queues rather than rejects data calls that cannot enter service immediately upon arrival. Another model extension is considered in Section 4.7, which presents a numerical sensitivity analysis of the SHL model of Chapter 3, differentiating between a high- and low-priority data service class. Section 4.8 ends this chapter with some concluding remarks.

4.1. LITERATURE

The relevant literature can be categorised according to the principal aspects of the investigated model, i.e. the PS service discipline, the service integration and the heavy tail of the service requirement distributions. We refer to Section 2.1 for a general literature review on the former two model aspects, stressing that a principal characteristic that is common to the reviewed integrated services analyses, is the considered Markovian model, in particular assuming exponentially distributed data call sizes. In this section we confine ourselves to a review of the most relevant investigations on the applicability and the impact of heavy-tailed service requirement distributions.

Since the seminal study [143] that first observed *self-similarity* and *long-range dependence* [183, 214] of traffic in Local Area Networks, a significant amount of research has concentrated on the observation, possible explication and performance implications of these phenomena. A stochastic process is called self-similar if it is scale-invariant in the sense that it shows similar behaviour over different time scales under appropriate scaling, and in particular exhibits a similar degree of burstiness across a range of time scales. A stochastic process is termed long-range dependent if its associated autocorrelation function is not integrable, and in particular if the autocorrelation function shows polynomial decay, which concerns its shape rather than its scale-invariance [183]. Although the notions of self-similarity and long-range dependence are not equivalent, and may indeed occur separately [232], mounting statistical evidence now exists that network traffic is characterised by both phenomena, as further indicated by the measurement reports for wide area TCP/IP networks [187] and for WWW traffic [56].

Aside from the 'single-source causality' for the occurrence of self-similarity and long-range dependence in telecommunication networks, that e.g. individual variable bit rate MPEG video streams may intrinsically exhibit long-range dependence [18, 117], the roots of a more structural causality can be attributed to the empirical property that the transfer of documents with a *heavy-tailed file size distribution* induces the aggregate network traffic to be long-range dependent at the multiplexing points. An analytical explanation for this causality relies on a superposition of a large number of independent on/off sources with heavy-tailed on- and/or off-periods, which constitutes a long-range dependent aggregate process [143, 223].

The impact of heavy-tailed file size distributions on *performance aspects* in PS-based systems has been (often asymptotically) analysed in e.g. [8, 118, 173, 233]. With regard to the sojourn time distribution in a PS queue under non-exponential service requirement distributions, [233] proved that for a regularly varying service requirement distribution (e.g. Pareto), the tails of the service requirement and sojourn time distributions are equally heavy. The generally considered desirable property of this tail equivalence is generalised in [173, Theorem 5.3.1] to an on/off server model (varying service capacity) and for the broader class of intermediately regularly varying service requirement distributions. Recently, the tail equivalence in the fixed capacity model has been extended to an even broader class of subexponential service requirement distributions, including e.g. lognormal and Weibull distributions (with shape parameter $\alpha < \frac{1}{2}$) [118]. As a comparison, note that for the $GI/G/1/FIFO$ queue it was shown in [53] that the tail of the sojourn time distribution is 'one degree' heavier than that of the service requirement distribution, which is due to the potentially severe impact of large calls queued ahead of small calls caused by the FIFO service discipline.

The influence of an adaptive flow control mechanism such as TCP on the performance is considered in [8]. The comparison presented in [8] demonstrates that whereas dramatic queue build-ups are indeed typical for autonomous long-range dependent processes without flow control, inclusion of TCP establishes dependencies among the different flows and leads to significantly more modest queue occupancies. In fact, it is observed that under TCP flow control light-tailed file size distributions cause more congestion than more heavy-tailed ones. These results thus strongly argue for a focus on *flow* level models (e.g. of PS-type) for performance evaluations of TCP/IP-based networks that include the considerable influence of flow control.

Another identified influence of heavy-tailed file size distributions on network performance, is that the induced long-range dependence enhances predictability of a call's continued presence, which can be exploited to devise efficient Call Admission Control rules or load balancing schemes [161, 183].

4.2. CONTRIBUTION

As a principal contribution of this chapter, the remarkable phenomenon is observed and analytically supported, that a greater call size variability leads to smaller (conditional) expected sojourn times in a Processor Sharing system with varying service capacity. Although the presented study is formulated in terms of and numerically analysed for the specific context of a single cell in an integrated services GSM/GPRS network, we stress that the qualitative conclusions are readily extended to a broader scope, e.g. a bottleneck link in an IP-based network, a cellular UMTS network or a WLAN. The investigated phenomenon is counterintuitive in view of previously published results for FIFO queues (e.g. [213]), where the reverse effect holds, or for fixed capacity PS queues [203, 204], where the expected sojourn times are insensitive to the call size variability. In contrast with such fixed capacity queueing models, delay-tolerant data services typically experience a varying capacity in modern integrated services networks, which makes the considered model particularly topical. Aside from the higher resource efficiency that can be achieved by exploiting the enhanced system predictability [161, 183], our results thus indicate another unexpected benefit from the observed large call size variability in existing communication networks [56, 143].

4.3. MODEL

Consider a single cell in an integrated GSM/ GPRS network with C traffic channels that are shared between speech and data calls, with preemptive priority for speech calls (see Figure 4.1). The defining characteristics of the speech and data services are similar to those considered in Chapter 3 and briefly summarised below. Subsequently, the assumed call handling schemes are described, as well as the relevant performance measures.

4.3.1. TRAFFIC MODEL

The integrated services model includes speech and data services.

SPEECH SERVICE: Speech calls arrive according to a Poisson process with arrival intensity λ_{speech}. The duration of a speech call is assumed to have PDF φ_{speech} with mean μ_{speech}^{-1}, and the speech traffic load is denoted $\rho_{\text{speech}} \equiv$

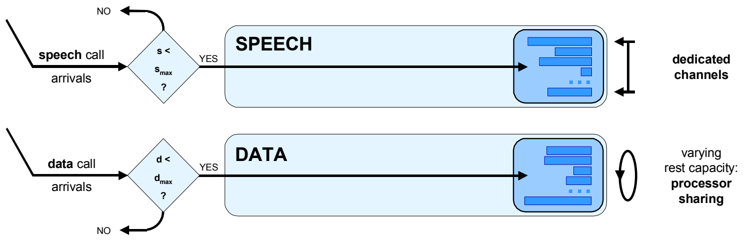


Figure 4.1 The considered GSM/GPRS system model integrates prioritised speech and delay-tolerant data calls sharing a common channel pool.

$\lambda_{\text{speech}}/\mu_{\text{speech}}$ (in Erlang). By definition, an admitted speech call is served with a fixed assignment of one traffic channel for its entire duration.

DATA SERVICE: Data calls arriving according to a Poisson process with arrival intensity λ_{data}, and comprise of a single file transfer. The *nominal* sojourn time (or: *size*) of a data call is defined as the call sojourn time under a fixed assignment of a single traffic channel and is assumed to have PDF φ_{data} with mean μ_{data}^{-1}. The involved normalisation is made with respect to the effective data bit rate per traffic channel r_{data} (in kbits/s), hence the mean file size of $1/\mu_{\text{data}}$ corresponds with an actual transfer volume of $r_{\text{data}}/\mu_{\text{data}}$ kbits. The data traffic load is denoted $\rho_{\text{data}} \equiv \lambda_{\text{data}}/\mu_{\text{data}}$ (in Erlang). The defining characteristic of a data call is that it is delay-tolerant and can handle a channel assignment that may vary between 0 and C. As a result, the actual sojourn time and the nominal sojourn time of a data call may differ. As a sequel to the numerical evaluation of the Chapter 3 SHL model, in Section 4.7 a distinction is made between a high- and low-priority data class.

4.3.2. SYSTEM MODEL

Denote with $S(t)$ and $D(t)$ the number of speech and data calls in the system at time t, respectively, with states denoted s and d. Define $\mathbb{S} \equiv \{0, 1, \cdots, C\}$. The call handling scheme serves speech calls with preemptive priority, so that an arriving speech call is blocked if and only if there are already $s_{\max} \equiv C$ speech calls present, in which case the call is cleared from the system. Admission of data calls is governed by a predetermined maximum number of data calls that is allowed in the system, denoted $d_{\max}$, and blocked data calls are cleared from the system. In contrast to the models studied in earlier chapters, as well as the extended model considered later in this chapter, in the basic model of Figure 4.1 no access queue is maintained to hold data calls that cannot be granted resources immediately upon arrival. At any time t, the channel sharing policy distributes the $C - S(t)$ available channels fairly over the present $D(t)$ data calls according to a PS service discipline without any per call data rate limitations. If at a given time all traffic channels are claimed by speech calls ($S(t) = C$), the transfer of any present data calls is preempted, while it is resumed upon the next speech call departure, at which time the capacity available to data traffic becomes non-zero again ($C - S(t) = 1 > 0$).

Section 4.6 considers an extension of the current model where data calls that cannot enter the PS queue immediately upon arrival are queued rather than blocked. Furthermore, in Section 4.7 a numerical sensitivity analysis is presented for a case of strict priority differentiation between two data service classes under a varying service capacity.

4.3.3. PERFORMANCE MEASURES

For speech calls the blocking probability $\mathbf{P}_{\text{speech}}$ is the only relevant performance measure, which is readily determined using the classical $M/G/C/C$ Erlang loss model. For data calls, we are primarily interested in the experienced QOS in terms of the expected call sojourn time $\mathbf{T}_{\text{data}}$ and the conditional expected call sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call of given size x. More specifically, as the document sizes of e.g. WWW calls are commonly acknowledged to be heavy-tailed [56, 143], we seek to determine the impact of the call size variability on these QOS measures. Although for finite $d_{\max}$ the data call blocking probability $\mathbf{P}_{\text{data}}$ is another essential performance measure that is considered, the principal result is most transparently conveyed for a model with $d_{\max} = \infty$, in order to prevent any ambiguity in the overall performance evaluation. This issue is revisited in the next section.

Since in a PS system with *fixed* capacity, the (conditional) expected sojourn time of a data call is independent of the call size distribution, we are also interested in the impact of the degree of variability in the service capacity on the QOS of data calls. To this end, the data service performance in the integrated services model is compared with that in a fixed capacity $M/G/1/d_{\max}/PS$ queueing model. Since this model is defined for a single channel, we must scale the data service requirements appropriately, resulting in $\mu^{\star}_{\text{data}} \equiv \mu_{\text{data}} C^{\star}$ and $\rho^{\star}_{\text{data}} \equiv \lambda_{\text{data}}/\mu^{\star}_{\text{data}} = \rho_{\text{data}}/C^{\star}$, with fixed channel capacity $C^\star \equiv C - \rho_{\text{speech}}(1 - \mathbf{P}_{\text{speech}})$, i.e. the average number of available channels in the integrated services model. The conditional expected sojourn time of a data call in the fixed capacity model is given by (see [54])

$$\mathbf{T}_{\text{data}}(x) = \frac{x}{C^\star} \left(\frac{\sum_{d=0}^{d_{\max}-1} \left(\rho^\star_{\text{data}}\right)^d (d+1)}{\sum_{d=0}^{d_{\max}-1} \left(\rho^\star_{\text{data}}\right)^d} \right),$$

while the expected data call sojourn time $\mathbf{T}_{\text{data}}$ trivially follows given that μ_{data}^{-1} is the expected data call size. Here $\mathbf{T}_{\text{data}}(x)$ and $\mathbf{T}_{\text{data}}$ are insensitive to the data call size distribution. For the specific case of $d_{\max} = \infty$, i.e. the standard $M/G/1/PS$ queueing model, we obtain $\mathbf{T}_{\text{data}}(x) = x/\left(C^\star - \rho_{\text{data}}\right)$ and $\mathbf{T}_{\text{data}} = 1/\left(\mu_{\text{data}}(C^\star - \rho_{\text{data}})\right)$ (see [213]), requiring $\rho^\star_{\text{data}} < 1$ for stability. As will be demonstrated, the $M/G/1/d_{\max}/PS$ model also serves as a limit case and lower bound for the (conditional) expected data call sojourn time in the integrated services model under extremely high data traffic load, regardless of the degree of variability in the service capacity.

4.4. OBSERVATIONS

An extensive simulation study has been carried out in order to investigate the impact of the data call size distribution tail and the degree of variability in the speech call arrival and termination process on the identified QOS measures.

Regarding the choice of the *speech* call duration distribution (φ_{speech}) a variety of options have been considered, including the exponential distribution and a bimodal mixture of lognormal distributions (see [49] for empirical results on the distribution of speech call durations in (cellular) communication networks). While the same qualitative results are demonstrated by all choices for the considered φ_{speech}, the presented graphs depict simulation results based on the exponentiality assumption, as it enables us to generate part of the results analytically (as in Chapters 2 and 3), and to provide analytical support in the next section. The included numerical results are however representative for the general trends.

Regarding the choice of the *data* call size distribution (φ_{data}), the simulation study comprises of a variety of options as well, including the Weibull and Pareto distributions, which have been selected to demonstrate the principal results. The specifications of the considered distributions are summarised in Table 4.1 below, where $\Gamma(\cdot)$ denotes the gamma function. Note that the deterministic PDF with $\varphi_{\text{data}}(x) = \delta\left(x - \mu_{\text{data}}^{-1}\right)$ is a special case of both the Weibull ($\alpha \to \infty$, $\beta = \mu_{\text{data}}^{-1}$) and Pareto ($\alpha \to \infty$, $c = \mu_{\text{data}}^{-1}$) PDFs, while the exponential PDF with $\varphi_{\text{data}}(x) = \mu_{\text{data}} e^{-x\mu_{\text{data}}}$ is a special case of the Weibull PDF ($\alpha = 1$ and $\beta = \mu_{\text{data}}^{-1}$).

Table 4.1 Considered data call size distributions: specifying parameters, probability density functions, means and coefficients of variation.

	WEIBULL, $\alpha, \beta > 0$	PARETO, $\alpha, c > 0$
$\varphi_{\text{data}}(x)$	$\alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}$, $x > 0$	$\alpha c^{\alpha} x^{-(\alpha+1)}$, $x > c$
μ_{data}^{-1}	$\frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$	$\begin{cases} \frac{\alpha c}{\alpha - 1} & \text{if } \alpha > 1 \\ \infty & \text{if } \alpha \leq 1 \end{cases}$
η_{data}	$\sqrt{2\alpha \frac{\Gamma\left(2\alpha^{-1}\right)}{\Gamma^{2}\left(\alpha^{-1}\right)} - 1}$	$\begin{cases} \frac{1}{\sqrt{\alpha(\alpha - 2)}} & \text{if } \alpha > 2 \\ \infty & \text{if } \alpha \leq 2 \end{cases}$

As a characterisation of the variability of a given data call size distribution, both the coefficient of variation η_{data} and the heaviness of the distribution tail, captured by shape parameter α, are considered. If it exists, η_{data} is an inversely proportional function of α, which is in correspondence with the property that for $x_0 > \beta$, $\Pr(x > x_0) = \exp\left(-\left(x_0/\beta\right)^{\alpha}\right)$ (Weibull) and for $x_0 > c$, $\Pr(x > x_0) = \left(c/x_0\right)^{\alpha}$ (Pareto) are decreasing in α, i.e. the tail becomes heavier as α becomes smaller. Note that the Pareto tail is generally heavier than the Weibull tail, regardless of the parameter choices, in the sense that

$$\lim_{x\to\infty} \frac{\int\limits_x^{\infty} \varphi_{\text{data}}^{weibull}(y)dy}{\int\limits_x^{\infty} \varphi_{\text{data}}^{pareto}(y)dy} = \lim_{x\to\infty} \frac{\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha_w}\right)}{\left(\frac{c}{x}\right)^{\alpha_p}} = 0,$$

for all α_w, β, α_p, $c > 0$, where the α-parameters of both distributions have been given an identifying subscript to express that they are generally different. Remark 4.1 at the end of the section comments on the applicability of these variability measures on the QOS comparison under different data call size distributions.

The Weibull distribution is particularly useful for our purposes as it enables a rather straightforward simulation study for a variety of η_{data}. In contrast, the Pareto distribution induces rather tedious simulations in order to obtain sufficient statistical accuracy. Still, the Pareto distribution is included in our study as it is probably the best-known distribution that satisfies all existing definitions of being truly *heavy-tailed*. In this light we refer to [232], where both the class of subexponential distributions (e.g. Weibull (for $\alpha \in (0,1)$), lognormal and Pareto) and the class of regularly varying distributions (generalisations of the Pareto distribution) are identified as heavy-tailed distributions. We stress that for our purposes it is not essential to consider heavy-tailed distributions per se, but rather to be able *to vary the balance between a small number of extremely large calls and a large number of small calls*, which is captured in the heaviness of the distribution tail, and study its impact on the experienced QOS.

As mentioned above, the parameter settings are based on the context of an integrated GSM/GPRS network. Typical values for such a case are $C = 22$, corresponding to 3 GSM frequencies, $\mu_{\text{speech}}^{-1} = 50$ seconds and $\rho_{\text{speech}} = 13.651$ Erlang, resulting in a speech call blocking probability of 1%, which is a typical target value for GSM operators and constant throughout the presented experiments. Regarding the data service, GPRS channel coding scheme CS-1 provides a per channel data rate of 9.05 kbits/s, so that the average nominal sojourn time of a data file with a mean size of 320 kbits, is $\mu_{\text{data}}^{-1} = 320/9.05 \approx 35.359$ seconds. The coefficient of variation η_{data} of the data calls is taken from the set $\left\{0, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16\right\}$ for the Weibull PDFs, while for the Pareto case we considered $\eta_{\text{data}} \in \left\{0, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, \infty_1, \infty_2\right\}$ to allow cross-PDF comparisons, where '∞_1' and '∞_2' denote two cases with infinite η_{data} ($\alpha_1 = 1.66$, $\alpha_2 = 1.35$, respectively [58], and the c's set to establish the intended average data call size) to present even heavier tails. EXPERIMENTS 1-3 assume $d_{\max} = \infty$, to prevent possible distortion of the results due to data call blocking, while EXPERIMENT 4 assumes a finite $d_{\max}$ in order to evaluate the system under an extreme data traffic load and to obtain insight into the distortion effect of data call blocking on the QOS.

In the figures included below to present the results of the four numerical experiments, each left chart is based on Weibull PDFs while the charts on the right represent Pareto PDFs. The 'o'-marked curves in the Weibull plots correspond to an exponential data call size distribution ($\eta_{\text{data}} = 1$), while the '•'-marked curves (all plots) correspond to the $M/G/1/PS$ queueing model with load $\rho_{\text{data}}^{\star}$, all of which can be calculated analytically. All other curves are obtained from simulations for which 95% confidence intervals have been determined with no worse than 5% relative precision (cases of finite η_{data} only). Since drawing samples from a data call size distribution with a large coefficient of variation, is non-trivial [32, 57], we have verified that the mean and variance (if this exists) of the sampled data call sizes correspond with the intended values.

Within each experiment, the marker for $\eta_{\text{data}} = 0$ as well as the $M/G/1/PS$ curves are identical in both the Weibull and Pareto plots. In order to help distinguish the different curves, we note that the included legends follow the order of the actual curves (top-to-bottom, with increasing η_{data}). In the charts' legends, '$\mathbf{T}_{\text{data}}$' is shortened to '$\mathbf{T}_{\text{d}}$', for reasons of readability. Regarding the applied units, we note that both the (conditional) expected sojourn times and the data call size (the nominal sojourn time) are expressed in *seconds*, the data traffic load is expressed in *Erlang*, and the call arrival rates are expressed in *calls per second*.

4.4.1. EXPERIMENT 1 (CONDITIONAL DATA QOS)

Figure 4.2 shows the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call as a function of its size x, for data traffic load $\rho_{\text{data}} = 6$, and for both the Weibull and the Pareto PDFs. For all $\eta_{\text{data}} > 0$, $\mathbf{T}_{\text{data}}(x)$ is displayed for data call sizes up to the 99%-percentile of the corresponding distribution, truncated at $x_{\max} = 200$ seconds in both plots in order to enable easy comparison (still capturing 97% (99%) of the most variable Weibull (Pareto) distribution). Whereas the Weibull distribution allows data call sizes down to 0 seconds, we note that the Pareto curves start at the corresponding c-values (minimum sample value), which is decreasing in η_{data}. Observe from the figures for both distribution types that *(i)* the greater the coefficient of variation of the data call size, the smaller the conditional expected sojourn times; equivalently, for the Pareto cases with infinite η_{data}: the smaller α, the heavier the tail and the smaller the conditional expected sojourn times; *(ii)* the curves all have a concave shape; *(iii)* conditional expected sojourn times are lowest in a system with fixed capacity. Furthermore, *(iv)* we observe that for a given $\eta_{\text{data}} > 0$ the conditional expected data call sojourn times appear to be lower for the Weibull distribution than for the corresponding Pareto distribution (see also Remark 4.1).

4.4.2. EXPERIMENT 2 (DATA QOS VERSUS TRAFFIC LOAD)

Figure 4.3 shows the expected sojourn time $\mathbf{T}_{\text{data}}$ of a data call as a function of the data traffic load ρ_{data} which is varied between 0 and 8 Erlang. The most important

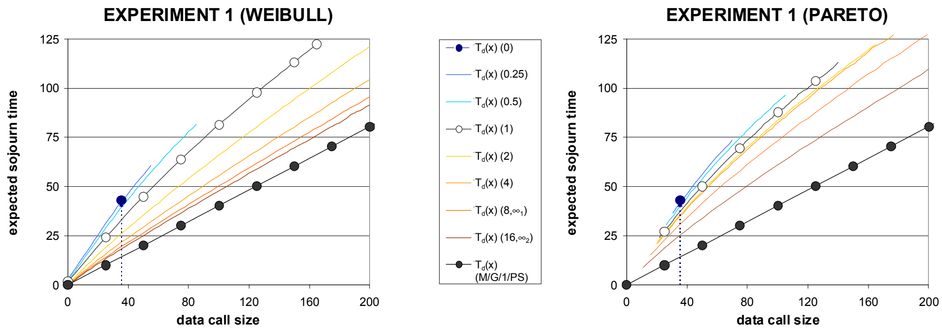


Figure 4.2 EXPERIMENT 1: Conditional expected data call sojourn times versus the data call size for different Weibull (left) and Pareto (right) data call size distributions.

observation that can be made from the figure is that *(i)* the greater the coefficient of variation of x, the smaller the expected sojourn time or equivalently, for the Pareto cases with infinite η_{data}, the smaller α, the heavier the distribution tail and the smaller the expected sojourn time. Observe further that *(ii)* for $\rho_{\text{data}} \to C^\star \approx 8.486$ Erlang, the expected sojourn times increase exponentially, whereas for even greater values of ρ_{data} the system becomes unstable. Note again that *(iii)* the data calls achieve optimal QOS in a system with fixed capacity, and that *(iv)* for a given η_{data} the QOS appears to be better for the Weibull distribution than for the corresponding Pareto distribution (see also Remark 4.1).

4.4.3. EXPERIMENT 3 (IMPACT OF SPEECH DYNAMICS)

In order to investigate the impact of the degree of variability in the service capacity on the QOS of data calls, the system has been simulated for the parameter settings of Figure 4.2, but varying λ_{speech} and μ_{speech} while keeping speech traffic load $\rho_{\text{speech}} = 13.651$ constant. The data traffic load is set to $\rho_{\text{data}} = 6$. The numerical results in Figure 4.4 allow a few conclusions. Most importantly, *(i)* the greater the rate of change of the service capacity, the smaller the difference in performance between the different data call size distributions. Intuitively this is not surprising as during the lifetime of a data call very rapid variations of the available capacity average out to a growing extent as λ_{speech} and μ_{speech} increase. In fact, the limiting scenario $(\lambda_{\text{speech}}, \mu_{\text{speech}} \to \infty)$ corresponds to a system with fixed capacity $C^\star \approx 8.486$ (fluid

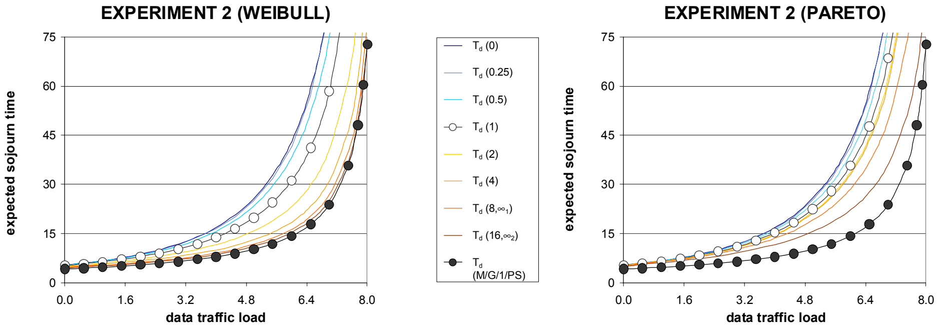


Figure 4.3 EXPERIMENT 2: Expected data call sojourn times versus the data traffic load for different Weibull (left) and Pareto (right) data call size distributions.

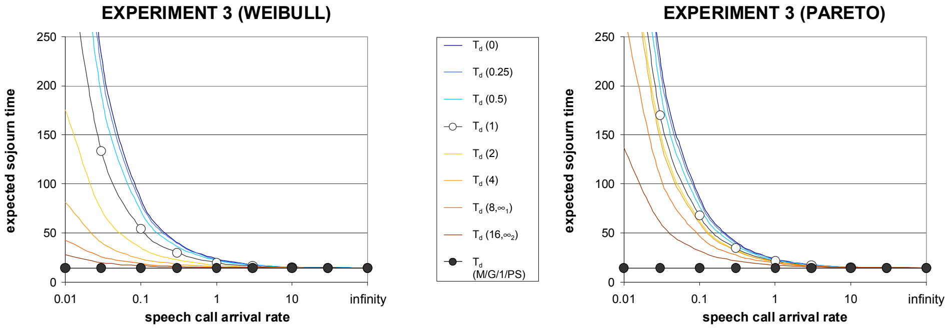


Figure 4.4 EXPERIMENT 3: Expected data call sojourn times versus the speech call arrival rate for different Weibull (left) and Pareto (right) data call size distributions.

limit). The expected data call sojourn time is then readily determined using standard results for an $M/G/1/PS$ queueing system: $\mathbf{T}_{\text{data}} = 1/\left(\mu_{\text{data}}(C^\star - \rho_{\text{data}})\right) \approx 14.226$. Furthermore, *(ii)* the figure provides additional support for the claim that the QOS experienced by data calls is better as the call sizes are more variable, as well as *(iii)* for the fact that the data QOS is best under fixed rather than varying capacity (see also [61]). Lastly, in accordance with the numerical results of EXPERIMENTS 1 and 2, *(iv)* for a given η_{data}, Weibull PDFs appear to yield better QOS than Pareto PDFs (see also Remark 4.1).

4.4.4. EXPERIMENT 4 (IMPACT OF DATA CAC)

As stated earlier, to prevent possible distortion of the results due to data call blocking, no data Call Admission Control (i.e. $d_{\max} = \infty$) was applied in order to obtain the above simulation results. Indeed, a final simulation experiment that has been carried out indicates that in case $d_{\max} < \infty$, data call blocking probability is lower if the data call sizes are more variable, implying a higher *carried* data traffic load, which in turn *may* yield higher data call sojourn times through increased competition for resources. The net effect of the data call size distribution is unclear in such a case, which is the very reason for avoiding data call blocking in the previous experiments. The numerically obtained expected sojourn times obtained for $d_{\max} = 100$ are plotted in Figure 4.5 for both the Weibull and the Pareto PDFs.

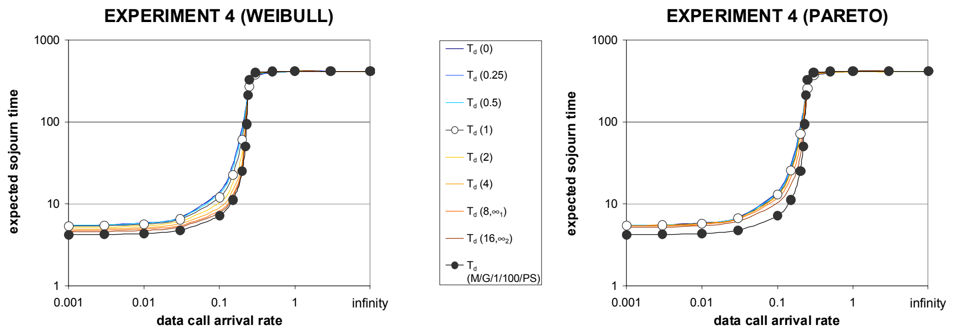


Figure 4.5 EXPERIMENT 4: Expected data call sojourn times versus the data call arrival rate for different Weibull (left) and Pareto (right) data call size distributions ($d_{\max} = 100$).

Three relevant observations can be made from the figure. The figure illustrates *(i)* the convergence of the expected data call sojourn times in a system with varying capacity to those in the corresponding $M/G/1/d_{\max}/PS$ queueing model. Furthermore, *(ii)* note that the expected sojourn times start to increase exponentially as $\rho_{\text{data}} \to C^{\star} \approx 8.486$ (cf. Figure 4.3), corresponding with $\lambda_{\text{data}} \to 0.240$, until the number of data calls present in the system becomes practically deterministic at $d_{\max}$ and the expected sojourn time flattens out at its maximum value. As an illustration of the above argument, *(iii)* note that for a range of λ_{data} values around 0.240 the data call QOS appears to be better (with 95% statistical significance) under a more light-tailed data call size distribution, which is due to a relatively strong heterogeneity in the carried data traffic load. In fact, for this range of λ_{data} the $M/G/1/d_{\max}/PS$ system with fixed capacity performs worst in QOS and best in GOS (data call blocking probability).

Although it was not so clear in Figure 4.3 due to the linearity of the vertical axis, Figure 4.5 (Weibull PDF) more clearly indicates that as $\lambda_{\text{data}} \to 0$, the expected data call sojourn times converge to distinct values for the different η_{data}, which are ordered in accordance with our principal result. Note that in such a case of extremely light data traffic, the data call size variability affects the experienced QOS of a tagged data call solely via the size distribution of the tagged data call itself, yet *not* via the sizes of other competing data calls as well. The reason for pointing out this extreme case is that it is one of the cases considered in the next section to provide analytical support and intuition.

Remark 4.1 The above experiments have demonstrated that in a Processor Sharing system with varying service capacity, the data call QOS improves as the distribution tail becomes heavier, i.e. as smaller calls become more frequent, while the rare large calls become larger. This effect has been observed within a given family of data call size distributions. Comparison of the numerical results for the Weibull and Pareto distributions also reveal that a more general statement regarding the impact of the data call size distribution on the data QOS is rather difficult to give, as it is not trivial to identify the variability measure that fully captures the essence of its impact on the QOS. The impact is not purely determined by the coefficient of variation ($\sim$ variance, second moment), since in that case the Weibull and Pareto curves would have to coincide for a given η_{data}. Nor is the impact purely determined by the heaviness of the distribution tail, as the Pareto tail has been shown to generally outweigh the Weibull tail, while it is not the case in the presented numerical experiments that all Weibull curves lie strictly above all Pareto curves. Hence the QOS impact is determined by the characteristics of the data call size distribution in a broader sense. Refer to Remark 4.3 in Section 4.5.1 for a further elaboration on the cross-PDF comparisons.

The observed phenomenon indicates that, since document or WWW page sizes are highly variable [56, 143], wrongly assuming e.g. deterministic or exponentially distributed data call sizes (as is typically done for reasons of analytical tractability) may lead to an *under*estimation of the experienced QOS (i.e. an *over*estimation of the expected sojourn times), which in turn can result in conservative and inefficient Call Admission Control schemes or network planning guidelines. By means of an illustrative example, Remark 4.2 below gives an indication of the bottom-line impact that disregarding the investigated phenomenon can have on e.g. the derivation of network planning guidelines.

Remark 4.2 Consider a GSM/GPRS network serving speech and data calls that is to cover an area of 41532 km^2 (the size of The Netherlands). Let the speech traffic load density be 0.100 Erlang/km^2 with an average speech call duration of 50 seconds, while the data traffic load density is equal to 0.050 Erlang/km^2 with an average data call size of 320 kbits, corresponding to an average nominal sojourn time of $9.05/320 \approx 35.359$ seconds, as also assumed above. The number of traffic channels per cell is equal to 22. For a given speech call blocking probability requirement of 1%, and a maximum tolerable value of the expected data call sojourn time of 20 seconds, a bisection search yields a maximally allowable (and thus optimal) cell radius of 6.204 (6.299) km under the assumption that data call sizes are deterministic (exponentially distributed). The given parameter settings cause the data QOS requirement to be the stringent one. These optimal cell radii require the deployment of 344 (334) cells (base stations) to cover the entire service area. In order to demonstrate the potential impact of a large call size variability, suppose that the call sizes are actually e.g. Weibullian with a coefficient of variation of e.g. 8 or 16, for which the readily evaluated fixed capacity $M/G/1/PS$ queueing model provides a very good approximation (see Figure 4.3 (left)). The optimal cell radius under the presumed correct 'heavy tails' assumption is then equal to 6.573 km, corresponding with only 306 base stations that need to be deployed. The bottom-line impact is thus that an incorrect assumption that the call sizes are deterministic or exponentially distributed, e.g. for reasons of analytical tractability, leads to an excessive network investment by 12.418% or 9.150%, respectively. It is noted that over-investments would be even more significant if the network planner would realise that its (e.g. Markovian) dimensioning models underestimate the data call size distribution tail, and falsely correct the anticipated error by increasing rather than decreasing the base station density, given the rather counterintuitive character of the observed phenomenon, e.g. in contrast with the tails' influence in a FIFO queueing model. Lastly, note that over-investments do lead to better-than-required QOS levels, although the expected sojourn times are not very sensitive to the data traffic load around $\mathbf{T}_{\text{data}} = 20$ seconds (see Figure 4.3 (left)).

4.5. ANALYTICAL SUPPORT

This section provides theoretical support for the observations presented in Section 4.4, by means of an analytical treatment of two extreme cases of the model defined in Section 4.3, as well as a discussion of the performance in between these extreme cases. Section 4.5.1 treats the limit case of $\lambda_{\text{data}} \to 0^+$, proving that a greater data call size variability indeed leads to better QOS. Subsequently, Section 4.5.2 shows that at the other extreme, i.e. $\lambda_{\text{data}} \to \infty$ (and assuming $d_{\max} < \infty$ to ensure stability), the data QOS becomes not only *insensitive* to the data call size distribution, but also equal to that achieved in an $M/G/1/d_{\max}/PS$ system with *fixed* capacity. Finally, in Section 4.5.3 a brief intuitive discussion is provided regarding the intermediate case of $\lambda_{\text{data}} \in (0, \infty)$.

4.5.1. LIMIT CASE: $\lambda_{\text{data}} \to 0^+$

Consider the case of exponentially distributed speech call durations. The presented analysis for the limit case $\lambda_{\text{data}} \to 0^+$, which is characterised by the presence of never more than a single data call, is broken up into two stages. First we determine a closed-form expression for the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call, indicating that it is concave in x. The applied analysis follows the same lines as the conditional data QOS analyses presented in Chapters 2 and 3. Subsequently, we show that for a given concave $\mathbf{T}_{\text{data}}(x)$ the expected sojourn time is decreasing in the coefficient of variation of the data call size.

DETERMINE $\mathbf{T}_{\text{data}}(x)$

Denote with $\tau_s(x)$ the random sojourn time of an admitted data call of size x, that finds s active speech calls in the system upon arrival, and let $\widehat{\tau}_s(x) \equiv \mathbf{E}\{\tau_s(x)\}$ be its expectation. Define the vector $\widehat{\boldsymbol{\tau}}_+(x) \equiv (\widehat{\tau}_s(x),\ s \in \mathbb{S}_+)$, with $\mathbb{S}_+ \equiv \{0, 1, \cdots, C-1\}$ the set of 'speech states' where the present data traffic is sharing an aggregate assignment of at least a single traffic channel. For the limit case of $\lambda_{\text{data}} \to 0^+$, the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call of size x is then equal to

$$\mathbf{T}_{\text{data}}(x) = \sum_{s \in \mathbb{S}} \pi(s) \widehat{\tau}_s(x), \tag{4.1}$$

where $\pi(s)$ is the $M/M/C/C$ equilibrium probability that there are s speech calls in the system. We stress that the probability that an *admitted* data call finds s speech calls upon arrival is *not* in general equal to $\pi(s)$, since the thinned arrival process of *admitted* data calls needs not be Poisson, as will be demonstrated in the next section.

In the case of $\lambda_{\text{data}} \to 0^+$, however, the data call blocking probability converges to zero and thus the arrival process of *admitted* data calls becomes Poisson, so that PASTA can be applied as in (4.1).

In order to obtain explicit expressions for $\widehat{\tau}_s(x)$, $s \in \mathbb{S}$, we note that for $\lambda_{\text{data}} \to 0^+$ the number of data calls in the system never exceeds 1, so that we may set $d_{\max} = 1$, without affecting the results.

Proposition 4.1 *Let $\mathcal{Q}$ be the infinitesimal generator of the $M/M/C-1/C-1$ Markov chain model with speech call arrival rate λ_{speech} and average speech call duration μ_{speech}^{-1}, and let $\boldsymbol{\pi}$ be the equilibrium distribution vector of the $M/M/C/C$ model with the same rates. Let $\boldsymbol{\pi}_+ \equiv (\pi(s),\ s \in \mathbb{S}_+)$. Denote with $\mathcal{B}_0 \equiv diag\,(C-s,\ s \in \mathbb{S}_+)$ the diagonal matrix containing the number of channels available for a data call in states $s \in \mathbb{S}_+$. Given $\mathbf{u} = \left(1, 1, \cdots, 1 + \lambda_{speech}/\left(\mu_{speech} C\right)\right) \in \mathbb{R}^C$, let $\boldsymbol{\gamma}$ be the unique solution to the system of linear equations*

$$\mathcal{Q}\boldsymbol{\gamma} = \frac{\mathcal{B}_0 \mathbf{1}}{\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1}} - \mathbf{u}, \tag{4.2}$$

$$\boldsymbol{\pi}_+ \mathcal{B}_0 \boldsymbol{\gamma} = 0. \tag{4.3}$$

where $\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1} = C - \rho_{speech}\left(1 - \mathbf{P}_{speech}\right)$ is the average number of channels available for data calls with $\mathbf{P}_{speech} = \pi(C)$. Then the conditional expected sojourn time $\widehat{\tau}_s(x)$ of a data call of size x entering the system in the presence of s speech calls, $s \in \mathbb{S}$, is given by the expressions

$$\widehat{\boldsymbol{\tau}}_+(x) = \frac{x}{\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1}} \mathbf{1} + \left[\mathcal{I} - \exp\left\{x \mathcal{B}_0^{-1} \mathcal{Q}\right\}\right] \boldsymbol{\gamma}, \tag{4.4}$$

$$\widehat{\tau}_C(x) = 1/\left(\mu_{speech} C\right) + \widehat{\tau}_{C-1}(x). \tag{4.5}$$

Proof Although the result is a special case of the more general results presented in e.g. Proposition 3.4, a dedicated yet brief proof is included both for reasons of clarity and to make our claim self-contained. Moreover, inclusion of a dedicated proof demonstrates that in contrast with earlier analyses, for the considered limit case the derived expressions are insensitive to the data call size distribution, as the specifics of the distribution are not used in the derivations.

We begin by noting that (4.5) indeed holds, since a data call that finds C speech calls present upon arrival is idle for an expected time $1/\left(\mu_{\text{speech}} C\right)$ until one of the speech calls terminates, after which the unaffected data call finds itself in the presence of $C-1$ speech calls. Equation (4.4) is proven using marginal analysis in an equivalent manner as in the proof of e.g. Proposition 3.4. Let $s \in \mathbb{S}_+$ and consider a time interval of length $\Delta > 0$, with Δ sufficiently small such that the data call cannot terminate within this time, i.e. $\Delta < x/C$. Conditioning on all the possible events occurring in this interval, we get for $\widehat{\tau}_s(x)$:

$$\begin{aligned}\widehat{\tau}_s(x) \quad &= \quad \Delta + \lambda_{\text{speech}}\,\Delta\,\widehat{\tau}_{s+1}(x - O(\Delta)) + s\,\mu_{\text{speech}}\,\Delta\,\widehat{\tau}_{s-1}(x - O(\Delta)) \\ &\qquad + \left(1 - \lambda_{\text{speech}}\,\Delta - s\,\mu_{\text{speech}}\,\Delta\right)\,\widehat{\tau}_s\left(x - (C-s)\Delta\right) + o\left(\Delta\right),\end{aligned}$$

where all available channels $(C - s)$ are assigned to the single (tagged) data call. Rearranging terms and letting $\Delta \downarrow 0$, we obtain the system of differential equations

$$(C-s)\,\frac{\partial \widehat{\tau}_s(x)}{\partial x} = 1 + \lambda_{\text{speech}}\widehat{\tau}_{s+1}(x) + s\mu_{\text{speech}}\widehat{\tau}_{s-1}(x) - \left(\lambda_{\text{speech}} + s\mu_{\text{speech}}\right)\widehat{\tau}_s(x),$$

for all $s \in \mathbb{S}_+$. Using (4.5) these expressions may equivalently be written in matrix notation:

$$\mathcal{B}_0\,\frac{\partial}{\partial x}\widehat{\boldsymbol{\tau}}_+(x) = \mathbf{u} + \mathcal{Q}\,\widehat{\boldsymbol{\tau}}_+(x). \tag{4.6}$$

The system is complemented with the initial condition $\widehat{\boldsymbol{\tau}}_+(0) = \mathbf{0}$ reflecting the fact that the sojourn time $\tau_s(0)$, $s \in \mathbb{S}_+$, of an 'empty' data call is zero, almost surely.

The existence and uniqueness of a solution $\widehat{\boldsymbol{\tau}}_+(x)$ for every initial vector, and of a vector $\boldsymbol{\gamma}$ that satisfies (4.2) and (4.3), are ensured by [51, Chapter 1, Section 8] and [213], respectively. The proof is then completed by verifying that the claimed unique solution indeed satisfies the derived system of differential equations (4.6) and the associated initial condition. □

The resulting expressions for $\widehat{\tau}_s(x)$ may be convex or concave in x, or neither convex nor concave, depending on the choices of C and $s \in \mathbb{S}$. We have derived explicit expressions for $\widehat{\tau}_s(x)$ for $C \in \{1, 2, 3, 4\}$ and $s \in \mathbb{S}$, and have observed that $\widehat{\tau}_s(x)$ is typically convex for small s, neither convex nor concave for medium s, and concave for large s. For $C = 1$ and $C = 2$ the expressions are sufficiently compact to be included below. For $C = 1$ we find

$$\begin{cases} \widehat{\tau}_0(x) = \left(\rho_{\text{speech}} + 1\right) x \\ \\ \widehat{\tau}_1(x) = \widehat{\tau}_0(x) + \mu_{\text{speech}}^{-1}, \end{cases}$$

while for $C = 2$ we find

$$\begin{cases} \widehat{\tau}_0(x) = \dfrac{\rho_{\text{speech}}^2 + 2\rho_{\text{speech}} + 2}{2\left(\rho_{\text{speech}} + 2\right)} x + \\ \qquad\qquad + \dfrac{\rho_{\text{speech}}\left(\rho_{\text{speech}} + 1\right)}{\mu_{\text{speech}}\left(\rho_{\text{speech}} + 2\right)^2} \left(\exp\left\{-\dfrac{1}{2} x \mu_{\text{speech}} \left(\rho_{\text{speech}} + 2\right)\right\} - 1\right), \\ \\ \widehat{\tau}_1(x) = \dfrac{\rho_{\text{speech}}^2 + 2\rho_{\text{speech}} + 2}{2\left(\rho_{\text{speech}} + 2\right)} x + \\ \qquad\qquad - \dfrac{2\left(\rho_{\text{speech}} + 1\right)}{\mu_{\text{speech}}\left(\rho_{\text{speech}} + 2\right)^2} \left(\exp\left\{-\dfrac{1}{2} x \mu_{\text{speech}} \left(\rho_{\text{speech}} + 2\right)\right\} - 1\right), \\ \\ \widehat{\tau}_2(x) = \widehat{\tau}_1(x) + \left(2\mu_{\text{speech}}\right)^{-1}. \end{cases}$$

Note that for $C = 1$ both $\widehat{\tau}_0(x)$ and $\widehat{\tau}_1(x)$ are linear in x, whereas for $C = 2$, $\widehat{\tau}_0(x)$ is strictly convex while $\widehat{\tau}_1(x)$ and $\widehat{\tau}_2(x)$ are strictly concave. An expression for $\mathbf{T}_{\text{data}}(x)$ is then readily derived using (4.1) and the $M/M/C/C$ equilibrium distribution $\pi(s) = \mathbf{G}^{-1}\rho_{\text{speech}}^s/s!$, $s \in \mathbb{S}$, with $\mathbf{G} \equiv \sum_{s\in\mathbb{S}} \rho_{\text{speech}}^s/s!$ the appropriate normalisation constant. For $C = 1$, the conditional expected sojourn time is given by

$$\mathbf{T}_{\text{data}}(x) = \frac{\rho_{\text{speech}}}{\mu_{\text{speech}}\left(\rho_{\text{speech}} + 1\right)} + \left(\rho_{\text{speech}} + 1\right) x,$$

which is linear in x, while for $C = 2$ we obtain

$$\mathbf{T}_{\text{data}}(x) = \frac{\rho_{\text{speech}}^2}{2\mu_{\text{speech}}\left(\rho_{\text{speech}}^2 + 2\rho_{\text{speech}} + 2\right)} + \frac{\rho_{\text{speech}}^2 + 2\rho_{\text{speech}} + 2}{2\left(\rho_{\text{speech}} + 2\right)} x +$$

$$- \left(\frac{2\rho_{\text{speech}} \left(\rho_{\text{speech}} + 1\right)^2}{\mu_{\text{speech}} \left(\rho_{\text{speech}} + 2\right)^2 \left(\rho_{\text{speech}}^2 + 2\rho_{\text{speech}} + 2\right)} \right) \times$$
$$\times \left(\exp\left\{ -\frac{1}{2} x \mu_{\text{speech}} \left(\rho_{\text{speech}} + 2\right) \right\} - 1 \right),$$

which is strictly concave in x, as are the equivalent expressions for $C \in \{3, 4\}$.

The expressions for the data call sojourn time $\mathbf{T}_{\text{data}}(x)$ have a straightforward interpretation that is most transparent for the case of $C = 1$. It is readily verified that $\mathbf{T}_{\text{data}}(x)$ is equal to the nominal sojourn time (x) *plus* the expected residual waiting time upon arrival $(\pi(1)\, \mu_{\text{speech}}^{-1})$ *plus* the expected number of transfer interruptions $(\lambda_{\text{speech}}\, x)$ *times* the duration of such an interruption $(\mu_{\text{speech}}^{-1})$, applying Wald's equation (see e.g. [213]). To see that the expected number of transfer interruptions is given by $\lambda_{\text{speech}}\, x$, note that it is equal to the expected number of Poisson arrivals (at rate λ_{speech}) during the nominal sojourn time x.

A GREATER DATA CALL SIZE VARIABILITY IMPROVES THE QOS

In this subsection we demonstrate that a greater data call size variability improves the QOS if the conditional expected data call sojourn time $\mathbf{T}_{\text{data}}(x)$ is concave in x, starting with a simple yet very insightful intuitive argument. Although in general $\mathbf{T}_{\text{data}}(x)$ may depend on the data call size distribution, it is insensitive in the considered limit case of $\lambda_{\text{data}} \to 0^+$, as is shown above. This enables us to compare the effect of the data call size distribution φ_{data} on the expected sojourn time $\mathbf{T}_{\text{data}}$ for a single given $\mathbf{T}_{\text{data}}(x)$.

Starting with a degenerate distribution in $x = \mu_{\text{data}}^{-1}$ with a coefficient of variation equal to zero and $\mathbf{T}_{\text{data}} = \mathbf{T}_{\text{data}}\left(\mu_{\text{data}}^{-1}\right)$, we increase the coefficient of variation by shifting an equal amount of probability mass equally far up and down the x scale, such that the mean data call size remains μ_{data}^{-1} (see Figure 4.6). Due to the concavity of $\mathbf{T}_{\text{data}}(x)$ the QOS gain that is made by moving some probability mass towards the lower end of the x scale, outweighs the QOS cost that is incurred by moving the same probability mass equally far towards the upper end of the x scale, so that indeed the net effect is a QOS improvement.

After this rather intuitive argument, two more rigorous approaches follow to support the claim that if the expected sojourn time $\mathbf{T}_{\text{data}}(x)$ is concave in x the expected

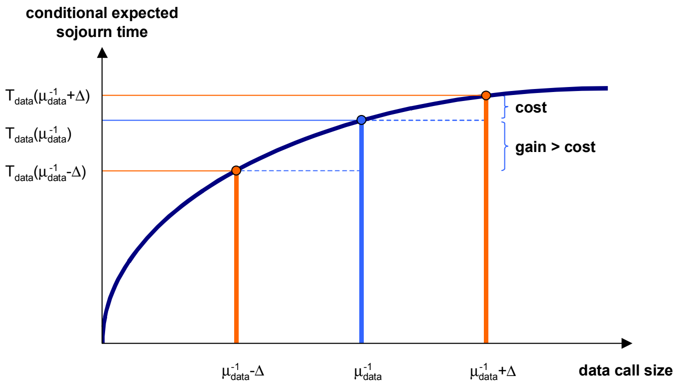


Figure 4.6 Analytical support: under a concave $\mathbf{T}_{\text{data}}(x)$, a greater data call size variability improves the QOS.

sojourn time $\mathbf{T}_{\text{data}}$ is lower if the PDF φ_{data} has a larger coefficient of variation (is more variable), given the same mean. Firstly, for the case that the data call size has a discrete PDF on two values, Proposition 4.2 below proves analytically that concavity of $\mathbf{T}_{\text{data}}(x)$ implies that $\mathbf{T}_{\text{data}}$ is decreasing in η_{data}.

Proposition 4.2 *Assume that the data call size has a discrete* PDF *that can take on two values. If the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call of size x is concave in x, then the expected sojourn time $\mathbf{T}_{\text{data}}$ is decreasing in the coefficient of variation of the data call size, for a given mean $\mu_{\text{data}}^{-1} \equiv \mathbf{E}\{x\}$.*

Proof Denote the two possible data call sizes with $x_0 \equiv \mu_{\text{data}}^{-1} - \vartheta_0$ and $x_1 \equiv \mu_{\text{data}}^{-1} + \vartheta_1$, for $\vartheta_0 \in \left[0, \mu_{\text{data}}^{-1}\right]$ and $\vartheta_1 \in [0, \infty)$. It is readily verified that $\xi_0 \equiv \Pr\{x = x_0\} = \vartheta_1 / (\vartheta_0 + \vartheta_1)$ and $\xi_1 \equiv \Pr\{x = x_1\} = \vartheta_0 / (\vartheta_0 + \vartheta_1)$ must hold, in order to establish that $\mathbf{E}\{x\} = \mu_{\text{data}}^{-1}$. The coefficient of variation η_{data} is equal to $\mu_{\text{data}}\sqrt{\vartheta_0\vartheta_1}$, which is strictly increasing in both ϑ_0 and ϑ_1. For a given μ_{data}, the expected sojourn time of a data call can be written as a function of ϑ_0 and the variance $\sigma_{\text{data}}^2 \equiv \vartheta_0\vartheta_1$:

$$\begin{aligned}\theta\left(\vartheta_0, \sigma_{\text{data}}^2\right) &= \frac{\sigma_{\text{data}}^2/\vartheta_0}{\vartheta_0 + \sigma_{\text{data}}^2/\vartheta_0}\mathbf{T}_{\text{data}}\left(\mu_{\text{data}}^{-1} - \vartheta_0\right) + \\ &\quad + \frac{\vartheta_0}{\vartheta_0 + \sigma_{\text{data}}^2/\vartheta_0}\mathbf{T}_{\text{data}}\left(\mu_{\text{data}}^{-1} + \sigma_{\text{data}}^2/\vartheta_0\right),\end{aligned}$$

where $\theta(\cdot,\cdot)$ denotes the expected sojourn time $\mathbf{T}_{\text{data}}$ which is a function of ϑ_0 and σ^2_{data}. In order to prove the proposition, it suffices to show that $\theta\left(\vartheta_0,\sigma^2_{\text{data}}\right)$ is decreasing in σ^2_{data}, since the coefficient of variation η_{data} is a simple increasing function of the variance σ^2_{data}. The first derivative of $\theta\left(\vartheta_0,\sigma^2_{\text{data}}\right)$ with respect to σ^2_{data} is given by

$$\begin{aligned}\frac{\partial}{\partial\sigma^2_{\text{data}}}\theta\left(\vartheta_0,\sigma^2_{\text{data}}\right) &= \frac{\mathbf{T}'_{\text{data}}\left(\mu^{-1}_{\text{data}}+\sigma^2_{\text{data}}/\vartheta_0\right)}{\vartheta_0+\sigma^2_{\text{data}}/\vartheta_0}+\\ &\quad -\frac{\mathbf{T}_{\text{data}}\left(\mu^{-1}_{\text{data}}+\sigma^2_{\text{data}}/\vartheta_0\right)-\mathbf{T}_{\text{data}}\left(\mu^{-1}_{\text{data}}-\vartheta_0\right)}{\left(\vartheta_0+\sigma^2_{\text{data}}/\vartheta_0\right)^2}.\end{aligned}$$

Using the property of concave functions that the tangent in any point lies above the function itself, the result immediately follows:

$$\left(\vartheta_0+\sigma^2_{\text{data}}/\vartheta_0\right)\mathbf{T}'_{\text{data}}\left(\mu^{-1}_{\text{data}}+\sigma^2_{\text{data}}/\vartheta_0\right)\leq\mathbf{T}_{\text{data}}\left(\mu^{-1}_{\text{data}}+\sigma^2_{\text{data}}/\vartheta_0\right)+$$

$$-\mathbf{T}_{\text{data}}\left(\mu^{-1}_{\text{data}}-\vartheta_0\right)\Leftrightarrow\frac{\partial}{\partial\sigma^2_{\text{data}}}\theta\left(\vartheta_0,\sigma^2_{\text{data}}\right)\leq 0.$$

□

Secondly, for the Weibull and Pareto PDFs introduced in Section 4.4 (including the deterministic case), Table 4.2 contains the expected data call sojourn times,

$$\mathbf{T}_{\text{data}}\equiv\int_{x=0}^{\infty}\mathbf{T}_{\text{data}}(x)\ \varphi_{\text{data}}(x)\,dx,$$

given $C\in\{1,2,3,4\}$ and the conditional data call sojourn times $\mathbf{T}_{\text{data}}(x)$ derived above. Recall that the two Pareto cases with infinite coefficient of variation are defined by $\alpha_1=1.66$, $\alpha_2=1.35$, and the c's such that the expected value is as intended. The speech call process is defined by an average call duration of $\mu^{-1}_{\text{speech}}=50$ seconds, while for each C the speech call arrival rate λ_{speech} is set such that the resulting speech traffic load $\rho_{\text{speech}}\equiv\lambda_{\text{speech}}/\mu_{\text{speech}}$ induces a 1% speech call blocking probability.

Since for $C=1$, $\mathbf{T}_{\text{data}}(x)$ is linear in x, the shape of the data call size distribution has no impact on the expected sojourn times. For $C\in\{2,3,4\}$, i.e. the cases where $\mathbf{T}_{\text{data}}(x)$ is strictly concave in x, the values in the table demonstrate that indeed

Table 4.2 Limit case that $\lambda_{\text{data}} \to 0^+$: expected data call sojourn times for various C and η_{data}, for both the Weibull and Pareto PDFs.

WEIBULL

C	ρ_{speech}				η_{data}					$M/G/1/PS$	
		0	1/4	1/2	1	2	4	8	16	($\star$)	($\diamond$)
1	0.010	36.22	36.22	36.22	36.22	36.22	36.22	36.22	36.22	36.22	35.72
2	0.153	20.38	20.36	20.31	20.19	19.98	19.78	19.64	19.55	19.37	19.12
3	0.455	15.30	15.28	15.23	15.09	14.85	14.60	14.41	14.29	14.04	13.87
4	0.869	12.73	12.71	12.67	12.53	12.28	12.02	11.81	11.67	11.39	11.26

PARETO

C	ρ_{speech}				η_{data}					$M/G/1/PS$	
		0	1/4	1/2	1	2	4	∞_1	∞_2	($\star$)	($\diamond$)
1	0.010	36.22	36.22	36.22	36.22	36.22	36.22	36.22	36.22	36.22	35.72
2	0.153	20.38	20.36	20.34	20.30	20.27	20.26	20.19	20.05	19.37	19.12
3	0.455	15.30	15.28	15.26	15.22	15.18	15.17	15.08	14.91	14.04	13.87
4	0.869	12.73	12.72	12.69	12.65	12.62	12.61	12.51	12.33	11.39	11.26

the expected data call sojourn time is decreasing in η_{data}. For the Pareto cases with $\eta_{\text{data}} = \infty$, the QOS improves with lower α (heavier tail). The final column of Table 4.2 (marked ($\diamond$)) contains the expected sojourn times $\mathbf{T}_{\text{data}} = 1/\mu^{\star}_{\text{data}}$ of a data call if it were served in a fixed capacity $M/G/1/PS$ system (with $\lambda_{\text{data}} \to 0^+$), while in the adjacent column (marked ($\star$)) these values are raised by $\pi(C)/(\mu_{\text{speech}}C)$, the expected delay before an admitted data call may start its transfer in the system with varying capacity. The significance of the latter values is that they are limit values for the system with varying capacity, as the data call size variability grows. Another observation that can be made from the table, is that the QOS improves as C increases, which is trivial as more capacity remains to be assigned to each admitted data call. Finally, for a given $\eta_{\text{data}} < \infty$, the QOS induced by the Weibull data call size distribution is better than that induced by the Pareto distribution (see also Remark 4.1 and Remark 4.3 below).

Remark 4.3 As a sequel to Remark 4.1, we utilise the exact numerical results of Table 4.2 to provide additional insight in cross-PDF comparisons of the impact of the call size variability on the experienced QOS. As observed in e.g. EXPERIMENTS 1-3 in Section 4.4, the Weibull data call size distribution appears to yield better

QOS than the Pareto distribution, for the same mean and coefficient of variation. Consider e.g. the case of $\eta_{\text{data}} = 1$ in Table 4.2. In view of the intuitive argument that the QOS is influenced by the balance between a small number of extremely large calls and a large number of small calls, refer to Figure 4.7 for the corresponding Cumulative Distribution Functions (CDFs). In contrast to the Weibull CDF, which exhibits a significant probability mass for small data call sizes, the Pareto CDF is positive only for call sizes greater than $c \approx 20.713$, which is rather large compared to the mean. These distribution characteristics are in line with the intuitive argument given above. If we were to increase the call size variability (only) of the Pareto CDF, e.g. consider the cases with $\eta_{\text{data}} = \infty$, we learn that the QOS is equivalent (Table 4.2: '∞_1') or better ('∞_2') compared to the Weibullian case, in correspondence with the shift of probability mass towards smaller call sizes that is indicated in Figure 4.7 (dashed curves: the minimum call size c decrease to 14.058 and 9.167, respectively). It is further noted from additional simulation experiments (not included) that the observed and intuitively supported ordering of the data QOS curves associated with the Weibull and Pareto distributions is preserved for different mean data call sizes (as in fact are the relative shapes of the CDFs). Although fundamental insight has been provided, it remains an open issue to determine a well-defined variability measure that captures the essential characteristics of a PDF to allow cross-PDF comparisons.

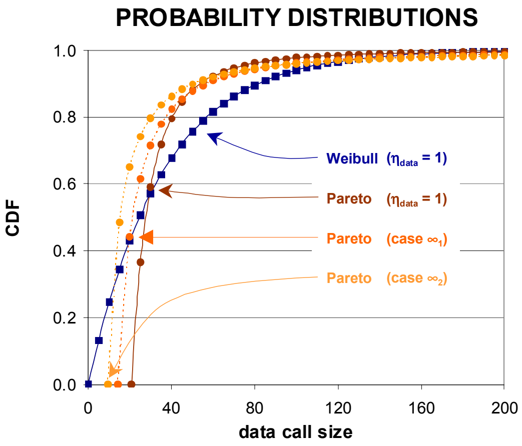


Figure 4.7 Cross-PDF comparisons: in comparison with Pareto PDFs with the same mean and variance, Weibull PDFs feature larger probability mass at relatively small data call sizes.

Remark 4.4 As a final elaboration on the matter of cross-PDF comparisons, following an alternate angle, the notion of one random variable being *stochastically more variable* than another random variable is a stochastic ordering principle that may be potentially applicable [199, Chapter 8, Section 5]. Consider two non-negative random variables X and Y, with $\mathbf{E}\{X\} = \mathbf{E}\{Y\}$. Then X is said to be stochastically more variable than Y, i.e. $X \geq_v Y$, if and only if $\mathbf{E}\{f(X)\} \geq \mathbf{E}\{f(Y)\}$, for all convex functions f. In order to understand this definition intuitively, take e.g. $f(a) = a^k$, $k \geq 1$, which is convex for all non-negative a. This choice of f indicates that a necessary requirement for X to be stochastically more variable than Y, is that all moments of X are at least as large as those of Y. Adapting this notion to our purposes here, the above definition is readily converted to state that if the conditional expected data call sojourn time $\mathbf{T}_{\text{data}}(x)$ is concave, and $X \geq_v Y$, then $\mathbf{E}\{\mathbf{T}_{\text{data}}(X)\} \leq \mathbf{E}\{\mathbf{T}_{\text{data}}(Y)\}$, which is in agreement with the observed phenomenon. It is however stressed that such a stochastic ordering is a sufficient, *not* necessary, condition to ensure the observed impact on the expected data QOS. Aside from the lack of a general proof of the suspected concavity of $\mathbf{T}_{\text{data}}(x)$, another problem is to verify the required conditions for the considered stochastic ordering for any two random variables. As proven in Proposition 4.2 (which is noted to exploit only the assumed concavity of $\mathbf{T}_{\text{data}}(x)$ and not its specific form), the considered stochastic ordering does apply to discrete distributions that take on two values. In a cross-PDF comparison of continuous Weibull and Pareto distributions with equal mean and coefficient of variation, however, it is readily shown that the stochastic ordering does not apply either way. For instance, the higher moments are generally larger for Pareto distributions, which is not surprising given their characteristically heavier tail. The k^{th} moment of a Pareto distribution is in fact infinite for all $k \geq \alpha$. Whereas this would seem to indicate that a Pareto random variable is stochastically more variable than a Weibull random variable with the same first and second moment, the numerical results based on the concave $\mathbf{T}_{\text{data}}(x)$ as derived above (see Table 4.2) show the opposite effect and thus serve as a counterexample. In conclusion, although the notion of 'stochastically more variable' may serve as a useful *sufficient* condition for a comparison of the experienced data QOS under different data call size distributions, the condition seems to be overly demanding.

ACCELERATED SPEECH CALL PROCESS

The results presented in Proposition 4.1 can be used to prove that as the speech call arrival and termination process is accelerated, i.e. $\lambda_{\text{speech}}, \mu_{\text{speech}} \to \infty$ while

ρ_{speech} remains fixed (fluid limit), the QOS converges to that of an $M/G/1/PS$ queueing system, and thus becomes insensitive to the data call size distribution. Define $\lambda^{\vartheta}_{\text{speech}} \equiv \vartheta\lambda_{\text{speech}}$ and $\mu^{\vartheta}_{\text{speech}} \equiv \vartheta\mu_{\text{speech}}$, with $\vartheta \in \mathbb{R}_{+}$, so that $\mathcal{Q}^{\vartheta} = \vartheta\mathcal{Q}$ is the infinitesimal generator of the Markov chain that describes the accelerated ($\vartheta > 1$) speech call process. The modification affects only $\mathcal{Q}$, so that the vector $\boldsymbol{\gamma}_{\vartheta}$ that solves the system (4.2), (4.3) is equal to $\boldsymbol{\gamma}/\vartheta$, with $\boldsymbol{\gamma}$ the solution in the basic model ($\vartheta = 1$). As a consequence, we find

$$\widehat{\boldsymbol{\tau}}_{+}(x) = \frac{x}{\boldsymbol{\pi}_{+}\,\mathcal{B}_0\,\mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{\vartheta x \mathcal{B}_0^{-1}\mathcal{Q}\right\}\right]\frac{\boldsymbol{\gamma}}{\vartheta} \to \frac{x}{\boldsymbol{\pi}_{+}\,\mathcal{B}_0\,\mathbf{1}}\mathbf{1} \quad (\vartheta \to \infty). \tag{4.7}$$

To see this, note that $\mathcal{B}_0^{-1}\mathcal{Q}$ is the infinitesimal generator of an irreducible finite state space Markov chain, with equilibrium distribution $\frac{\boldsymbol{\pi}\mathcal{B}_0}{\boldsymbol{\pi}\mathcal{B}_0\mathbf{1}}$, so that $\lim_{\vartheta\to\infty}\exp\left\{\vartheta x \mathcal{B}_0^{-1}\mathcal{Q}\right\} = \mathbf{1}\frac{\boldsymbol{\pi}\mathcal{B}_0}{\boldsymbol{\pi}\mathcal{B}_0\mathbf{1}}$. Also $\widehat{\tau}_C(x) \to \frac{x}{\boldsymbol{\pi}_{+}\mathcal{B}_0\mathbf{1}}$, as $\vartheta \to \infty$, which immediately follows from (4.5), (4.7) and the fact that $1/\left(\vartheta\mu_{\text{speech}}C\right) \to 0$, as $\vartheta \to \infty$. Hence

$$\mathbf{T}_{\text{data}}(x) = \sum_{s\in\mathbb{S}} \pi(s)\,\widehat{\tau}_s(x) \to \frac{x}{\boldsymbol{\pi}_{+}\,\mathcal{B}_0\,\mathbf{1}} \quad (\vartheta \to \infty),$$

which is precisely the conditional expected sojourn time in an $M/G/1/PS$ model with fixed capacity $\boldsymbol{\pi}_{+}\,\mathcal{B}_0\,\mathbf{1} = C^{\star} = C - \rho_{\text{speech}}\,(1 - \mathbf{P}_{\text{speech}})$ and an infinitesimally low (data) call arrival rate. It is obvious that then also the expected sojourn times must be equal.

4.5.2. LIMIT CASE: $\lambda_{\text{data}} \to \infty$

Consider the case of exponentially distributed speech call durations and assume $d_{\max} < \infty$. In this subsection we prove that in the limit of $\lambda_{\text{data}} \to \infty$, ensuring a continuous presence of $d_{\max}$ data calls, the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call is linear in x, insensitive to the data call size distribution, and moreover, $\mathbf{T}_{\text{data}}(x)$ is equal to the conditional expected sojourn time in the corresponding $M/G/1/d_{\max}/PS$ queueing model with fixed capacity.

DETERMINE $\mathbf{T}_{\text{data}}(x)$

With $\tau_s(x)$ and $\widehat{\tau}_s(x)$ as introduced in Section 4.5.1, define the vector $\widehat{\boldsymbol{\tau}}_{+}(x) \equiv (\widehat{\tau}_s(x),\ s \in \mathbb{S}_{+})$. The conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call of size x is then equal to

$$\mathbf{T}_{\text{data}}(x) = \sum_{s \in \mathbb{S}} \pi^{\star}(s)\, \widehat{\tau}_s(x),$$

where $\pi^{\star}(s)$ is the probability that an *admitted* data call finds s speech calls upon arrival, which is *not* equal to the equilibrium probability $\pi(s)$ that s speech calls are present in the system. For instance, $\pi(C) \neq \pi^{\star}(C) = 0$ since a data call must terminate in a system state $s \in \mathbb{S}_+$ while in the case of $\lambda_{\text{data}} \to \infty$ its freed service position is immediately taken by a fresh data call.

In order to determine the probabilities $\pi^{\star}(s)$, $s \in \mathbb{S}_+$, time is rescaled as illustrated by Figure 4.8 for the case of $C = 2$. The random time between two successive speech call arrival or termination events during which s speech calls are present in the system is weighted by the corresponding number of channels assigned to a data call, $d_{\max}^{-1}(C - s)$, so that on the new time scale a fixed capacity of one channel is continuously available for each of the data calls.

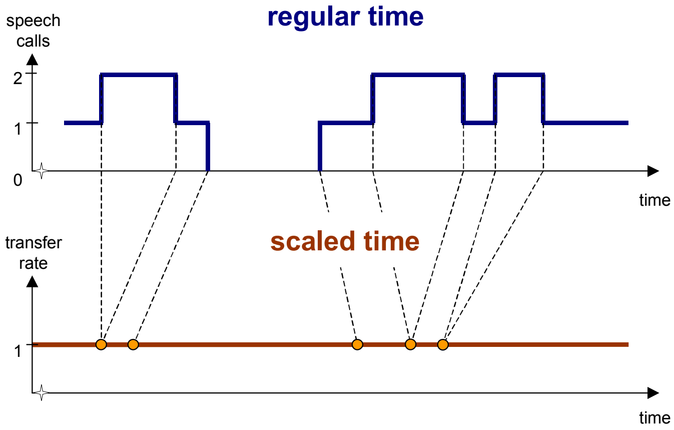


Figure 4.8 Analytical support: rescaling of time with respect to the varying channel availability for data traffic.

On the new time scale the coinciding data call arrival and termination instants form a renewal process with random interrenewal times equal to the data call sizes. The probability $\pi^{\star}(s)$ that a fresh data call finds s speech calls present upon admission is then given by the stationary distribution of the system state at such a renewal event, i.e. the fraction of time that the system serves s speech calls:

$$\pi^\star(s) \equiv \frac{d_{\max}^{-1}\,(C-s)\;\pi(s)}{\sum\limits_{s'\in\mathbb{S}_+} d_{\max}^{-1}\,(C-s')\;\pi(s')},\ s \in S_+$$
$$\Leftrightarrow \boldsymbol{\pi}_+^\star \equiv (\pi^\star(s),\ s \in \mathbf{S}_+) = \frac{\boldsymbol{\pi}_+\,\mathcal{B}_\infty}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}},$$

with $\pi(s)$ the time fraction that s speech calls are present in an $M/M/C/C$ Erlang loss model with speech traffic load ρ_{speech} on a regular time scale. The diagonal matrix $\mathcal{B}_\infty \equiv diag\left(d_{\max}^{-1}\,(C-s)\,, s \in \mathbb{S}_+\right) = d_{\max}^{-1}\mathcal{B}_0$ contains the number of channels available for an arbitrary data call in states $s \in \mathbb{S}_+$.

For the considered case of $\lambda_{\text{data}} \to \infty$, explicit expressions for $\widehat{\tau}_s(x)$, $s \in \mathbb{S}$, can be obtained by analogy with Proposition 4.1, using $\mathcal{B}_\infty$ rather than $\mathcal{B}_0$, as the number of data calls in the system is now continuously equal to $d_{\max}$. Note that the solution $\boldsymbol{\gamma}$ to the system of equations (4.2) and (4.3) is the same, regardless of whether $\mathcal{B}_0$ or $\mathcal{B}_\infty$ is used. In light of the earlier remark that an infinite data call arrival rate implies that no call is ever admitted in the presence of C speech calls and hence each data call can start service immediately upon admission, we note that nonetheless an expression is obtained for $\widehat{\tau}_C(x)$. Naturally, since $\pi^\star(C) = 0$, $\widehat{\tau}_C(x)$ does not contribute to $\mathbf{T}_{\text{data}}(x)$.

It is readily proven that for any C, $\mathbf{T}_{\text{data}}(x)$ is linear in x:

$$\begin{aligned}
\mathbf{T}_{\text{data}}(x) &= \frac{\boldsymbol{\pi}_+\mathcal{B}_\infty}{\boldsymbol{\pi}_+\mathcal{B}_\infty\mathbf{1}}\left\{\frac{x}{\boldsymbol{\pi}_+\mathcal{B}_\infty\mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{x\mathcal{B}_\infty^{-1}\mathcal{Q}\right\}\right]\boldsymbol{\gamma}\right\} \\
&= \frac{x}{\boldsymbol{\pi}_+\mathcal{B}_\infty\mathbf{1}} + \frac{\boldsymbol{\pi}_+\mathcal{B}_\infty}{\boldsymbol{\pi}_+\mathcal{B}_\infty\mathbf{1}}\left\{\sum_{i=1}^{\infty}\frac{\left(x\mathcal{B}_\infty^{-1}\mathcal{Q}\right)^i}{i!}\right\}\boldsymbol{\gamma} \\
&= \frac{x}{\boldsymbol{\pi}_+\mathcal{B}_\infty\mathbf{1}} \\
&= \frac{d_{\max}x}{C^\star}, \qquad (4.8)
\end{aligned}$$

using $\boldsymbol{\pi}_+\mathcal{Q} = 0$, due to the reversibility of the $M/M/C/C$ queue (note that $\boldsymbol{\pi}_+$ contains equilibrium probabilities of the $M/M/C/C$ queue, while $\mathcal{Q}$ is the infinitesimal generator of the $M/M/C-1/C-1$ queue). We stress that in the considered limit case expression (4.8) is *insensitive* to the specifics of the data call size distribution.

For $C = 1$, the conditional expected sojourn time is given by

$$\mathbf{T}_{\text{data}}(x) = d_{\max} \left(\rho_{\text{speech}} + 1\right) x,$$

which is simply equal to the linear function $\widehat{\tau}_0(x)$, while for $C = 2$ we obtain

$$\mathbf{T}_{\text{data}}(x) = d_{\max} \frac{\rho_{\text{speech}}^2 + 2\rho_{\text{speech}} + 2}{2\left(\rho_{\text{speech}} + 2\right)} x,$$

which is a weighted average of $\widehat{\tau}_0(x)$ and $\widehat{\tau}_1(x)$, whose respective convexity and concavity balance out, as is also the case for $C \in \{3, 4\}$.

THE QOS IS INSENSITIVE TO THE DATA CALL SIZE VARIABILITY

Since $\mathbf{T}_{\text{data}}(x)$ is linear in x for all C and finite $d_{\max}$, the shape of the data call size distribution has no impact on the expected sojourn times. Moreover, $\mathbf{T}_{\text{data}}$ is equal to $d_{\max}/\left(\mu_{\text{data}} C^{\star}\right)$, the average data call size divided by the average number of available channels. Note that this is precisely the expected sojourn time in an $M/G/1/d_{\max}/PS$ system:

$$\begin{aligned}\mathbf{T}_{\text{data}} &= \lim_{\lambda_{\text{data}} \to \infty} \frac{1}{\mu_{\text{data}}^{\star}} \left(\frac{\sum_{d=0}^{d_{\max}-1} \left(\rho_{\text{data}}^{\star}\right)^d (d+1)}{\sum_{d=0}^{d_{\max}-1} \left(\rho_{\text{data}}^{\star}\right)^d} \right) \\ &= \frac{d_{\max}}{\mu_{\text{data}} C^{\star}},\end{aligned}$$

using l'Hôpital's rule and appropriately scaling the data service requirements (as in Section 4.3.3). This result is not so surprising as for $\lambda_{\text{data}} \to \infty$ the distribution of the system state observed upon arrival by an admitted data call is independent of the data call size distribution, so that the fact that the system continuously contains $d_{\max}$ data calls implies that the expected amount of capacity an admitted data call enjoys is equal to the expected amount of capacity that is available, divided by the (deterministic) number $d_{\max}$ of data calls sharing this capacity.

ACCELERATED SPEECH CALL PROCESS

As a final note, in contrast with the results of Section 4.5.1, the (conditional) expected sojourn times are insensitive to an *acceleration* of the speech call arrival and termination process, since it influences $\mathbf{T}_{\text{data}}\left(x\right)$ only through ρ_{speech} and not through λ_{speech} and μ_{speech} individually.

4.5.3. INTERMEDIATE CASE: $\lambda_{\textbf{data}} \in (0, \infty)$

So far we have investigated the effect of the data call size distribution on the QOS measures for the extreme cases of $\lambda_{\text{data}} \to 0^+$ and $\lambda_{\text{data}} \to \infty$. The principal reason why these limit cases are analytically tractable is that the level of competition of a tagged data call is independent of the data call size distribution, as the number of data calls competing for the same resources is either 0 for $\lambda_{\text{data}} \to 0^+$, or $d_{\max} - 1$ for $\lambda_{\text{data}} \to \infty$. Moreover, in these cases the distribution of the number of present speech calls that an admitted data calls sees upon arrival can be explicitly determined and is insensitive to the data call size distribution. For $\lambda_{\text{data}} \in (0, \infty)$, however, the number of competing data calls that a tagged data call endures is not only a random variable, but it is also sensitive to the data call size distribution. As a consequence, the distribution of the system state upon departure of a data call is also sensitive to the data call size distribution, and hence the distribution of the system state upon admission of a data call is as well.

As an illustrative argument, recall from Section 4.5.1 the expressions for the conditional expected call sojourn time $\widehat{\tau}_s(x)$, $s \in \mathbb{S}$, of a data call of size x, admitted to the system in the presence of s speech calls (with $d_{\max} = 1$). It was noted that for $C = 2$, $\widehat{\tau}_0(x)$ is strictly convex while $\widehat{\tau}_1(x)$ and $\widehat{\tau}_2(x)$ are strictly concave. The weighted average $\mathbf{T}_{\text{data}}(x) \equiv \sum_{s \in \mathbb{S}} p(s)\,\widehat{\tau}_s(x)$ of these expressions, where $p\left(s\right)$ denotes the probability that an arbitrary data call finds s speech calls present upon admission, is concave if sufficient weight lies on the concave $\widehat{\tau}_1(x)$ and $\widehat{\tau}_2(x)$, i.e. if $p(0)$ is sufficiently small. We conjecture that $p(0)$ increases monotonously from $\pi(0)$ to $\pi^\star(0)$, as λ_{data} runs from 0 to ∞, which is supported by the results obtained from some additional simulations (not included). Besides the fact that this monotonicity implies that indeed $\mathbf{T}_{\text{data}}(x)$ is concave for all λ_{data}, so that a greater data call size variability enhances the QOS, the aforementioned additional simulation results further revealed that for $\lambda_{\text{data}} \in (0, \infty)$, the probability $p(0)$ decreases in the data call size variability, providing additional support for the observed trend (in view of the discussion in Section 4.4

regarding the potentially distorting effect of a finite $d_{\max}$ on data call blocking and the experienced QOS, we note that this phenomenon was not observed for the case of $d_{\max} = 1$ considered here for illustrative purposes).

4.6. AN EXTENDED MODEL

In the previous sections we have demonstrated and analytically supported the phenomenon that QOS improves under more variable data call sizes in a PS model with varying capacity. We would now like to shed some light on an interesting extension of the basic model studied in the previous sections. As opposed to the basic model of Figure 4.1, in the extended model (see Figure 4.9) data calls that cannot find a free position in the PS queue (*transfer queue*) are not blocked but rather queued in an infinite FIFO *access queue*. Recall that this extended model was also studied in Chapters 2 and 3 for exponentially distributed data call sizes, while [9] considers the data call performance in isolation, i.e. without the capacity fluctuations due to the prioritised speech call arrival and departure process.

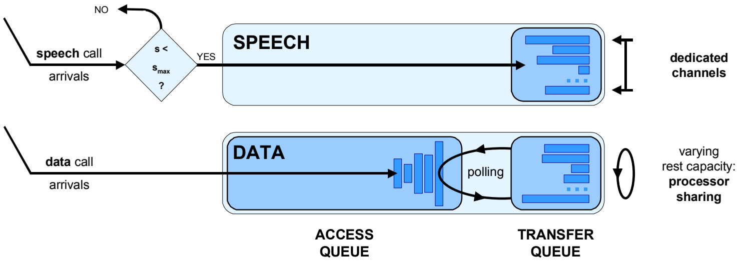


Figure 4.9 Compared to the basic system model, the extended system model maintains an access queue for data calls that cannot enter transfer immediately upon arrival.

Aside from the appealing additional insight that the evaluation of the extended model conveys, given the conflicting impact of the call size variability on the QOS of a stand-alone FIFO or PS queue (with varying service capacity), the extended model also resembles the practical operations of e.g. a GSM/GPRS access network. In such networks, a limit can be placed on the concurrent number of data transfers, while a so-called 'packet queueing notification' [74] can be sent to data call requests that cannot be honored immediately, effectively placing them in a FIFO queue. As the numerical results presented below indicate, the number of service positions in the transfer queue can have a significant impact on the experienced QOS and should therefore be carefully chosen.

The number of service positions in the transfer queue is denoted by $d_{\max}^{\text{transfer}}$, and we are primarily interested in the impact of $d_{\max}^{\text{transfer}}$ on the relative performance of the different data call size distribution tails. Moreover, we compare this impact in both the integrated services model with varying PS capacity and the model with fixed PS capacity $C^\star \equiv C - \rho_{\text{speech}}(1 - \mathbf{P}_{\text{speech}})$, i.e. the *average* number of available channels in the integrated services model.

With regard to the *fixed* capacity model, we note that the extreme cases of $d_{\max}^{\text{transfer}} \in \{1, \infty\}$ represent the pure FIFO and PS models, respectively, and allow exact expressions for the expected data call sojourn times. For the case $d_{\max}^{\text{transfer}} = 1$ (FIFO), the expected data call sojourn time is given by the well-known Pollaczek-Khintchine formula (see e.g. [213]):

$$\mathbf{T}_{\text{data}} = \frac{\left(\eta_{\text{data}}^2 + 1\right)}{2} \frac{\rho_{\text{data}}/C^\star}{\mu_{\text{data}}\left(C^\star - \rho_{\text{data}}\right)} + \frac{1}{\mu_{\text{data}} C^\star},$$

implying that heavier tails (higher η_{data}) induce *worse* QOS. At the other extreme of $d_{\max}^{\text{transfer}} = \infty$ (PS), the expected data call sojourn time is equal to

$$\mathbf{T}_{\text{data}} = \frac{1}{\mu_{\text{data}}\left(C^\star - \rho_{\text{data}}\right)},$$

as already stated in Section 4.3, which expresses insensitivity of $\mathbf{T}_{\text{data}}$ with respect to η_{data}. In [9] the following approximation of the expected data call sojourn time is presented for a fixed capacity system that applies for all $d_{\max}^{\text{transfer}}$:

$$\mathbf{T}_{\text{data}} \cong \frac{\left(\eta_{\text{data}}^2 + 1\right)}{2} \frac{\left(\rho_{\text{data}}/C^\star\right)^{d_{\max}^{\text{transfer}}}}{\mu_{\text{data}}\left(C^\star - \rho_{\text{data}}\right)} + \frac{1 - \left(\rho_{\text{data}}/C^\star\right)^{d_{\max}^{\text{transfer}}}}{\mu_{\text{data}}\left(C^\star - \rho_{\text{data}}\right)}, \tag{4.9}$$

where the first term approximates the expected access time and the second term approximates the expected transfer time. It is readily verified that the approximation provides exact results for $d_{\max}^{\text{transfer}} \in \{1, \infty\}$, representing pure FIFO and PS models, respectively, as well as for the case of exponentially distributed data call sizes, where $\mathbf{T}_{\text{data}}$ is independent of $d_{\max}^{\text{transfer}}$. Both the exact extreme cases and the approximation suggest that lower η_{data} yield better QOS for small $d^{\text{transfer}}_{\text{max}}$ while its impact vanishes as $d^{\text{transfer}}_{\text{max}}$ increases.

4.6.1. EXPERIMENT 5F (FIXED CAPACITY)

Figure 4.10 shows the numerical results that support this expectation for both the Weibull and the Pareto PDFs (EXPERIMENT 5F). The plotted values are obtained by exact calculations where possible and simulations elsewhere. It is noted that in the Pareto cases with $\eta_{\text{data}} = \infty$ the expected data call sojourn times are infinite for small $d^{\text{transfer}}_{\text{max}}$ and finite for large $d^{\text{transfer}}_{\text{max}}$, and simulation experiments as used to generate Figure 4.10 can only loosely indicate the minimum transfer queue size that guarantees a finite expected sojourn time. Observe that, as also shown in [204], the expected sojourn time in a pure FIFO model is lower (higher) than that in a pure PS model if the coefficient of variation of the data call size distribution is smaller (greater) than 1. As a side result, approximation (4.9) appears to be rather good for the Weibullian data call sizes but very poor for the Pareto case, especially for moderate values of $d^{\text{transfer}}_{\text{max}}$, where it occasionally even underestimates $\mathbf{T}_{\text{data}}$ by a factor greater than five.

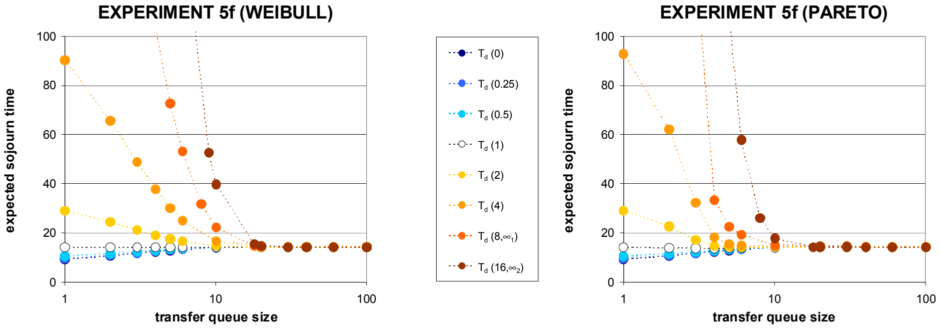


Figure 4.10 EXPERIMENT 5F: Expected data call sojourn times versus the transfer queue size for different Weibull (left) and Pareto (right) data call size distributions.

4.6.2. EXPERIMENT 5V (VARYING CAPACITY)

For the more interesting case of a *varying* server capacity we expect that for small values of $d^{\text{transfer}}_{\text{max}}$ the FIFO queue dominates and a heavier tail degrades the QOS.

We argue that the varying capacity does not affect the qualitative phenomenon that relatively small data calls suffer greatly from relatively large data calls ahead of them in the queue, which is typical for FIFO queues. On the other hand, as $d_{\max}^{\text{transfer}}$ increases, the significance of the FIFO access queue diminishes and the system performance is more and more determined by the PS transfer queue. Based on the observations and analysis in Sections 4.4 and 4.5 we know that the reverse impact of η_{data} on $\mathbf{T}_{\text{data}}$ then applies. The numerical results in Figure 4.11 demonstrates the expected reversal of the ordering of the $\mathbf{T}_{\text{data}}$ curves as $d_{\max}^{\text{transfer}}$ is raised from 1 to ∞. We claim that in a well-dimensioned network, the access time is relatively small compared to the transfer time, corresponding with a relatively large number of service positions $d_{\max}^{\text{transfer}}$ (with respect to the typical occupation of the access queue), so that the 'PS effect' dominates: the greater the data call size variability, the better the QOS.

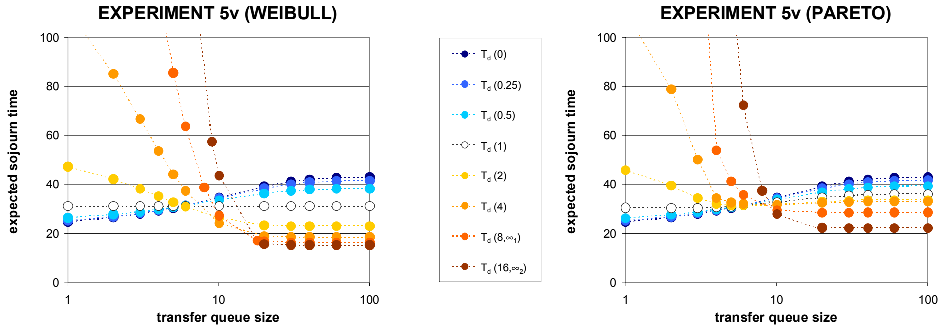


Figure 4.11 EXPERIMENT 5V: Expected data call sojourn times versus the transfer queue size for different Weibull (left) and Pareto (right) data call size distributions.

The above remark regarding the finiteness of the expected data call sojourn time of the Pareto cases with $\eta_{\text{data}} = \infty$ also applies here. Observe further that there is no generally uniform $d_{\max}^{\text{transfer}}$ where the ordering is reversed, although the curves corresponding to $\eta_{\text{data}} \leq 2$ do appear to jointly cross one another at about $d_{\max}^{\text{transfer}} = 6$ (both distributions).

Comparing the figures of EXPERIMENTS 5F and 5V, we observe that for all depicted cases the QOS under varying capacity is worse than under fixed capacity, while for small (large) $d_{\max}^{\text{transfer}}$ the absolute difference increases (decreases) in the data call size variability. In particular, in the extreme case of $d_{\max}^{\text{transfer}} = 1$ (FIFO) the absolute

QOS differences between the varying and fixed capacity scenarios worsens as the data call sizes become more variable, while in the extreme case of $d_{\max}^{\text{transfer}} = \infty$ (PS) the reverse effect is observed.

4.7. A MODEL WITH QOS DIFFERENTIATION

The final sensitivity analysis concentrates on the SHL model of Chapter 3, integrating speech, high- and low-priority data calls. Refer to Section 3.4.4 for a detailed description of the SHL model, which assumes a Discriminatory Processor Sharing discipline to share the varying resources among the different data classes.

4.7.1. EXPERIMENT 6 (DATA QOS DIFFERENTIATION)

In order to allow comparison with the earlier experiments in this chapter, we depart slightly from the default parameter settings assumed in Section 3.7. More specifically, we assume an infinite transfer queue, no multislot restrictions, no Radio Resource Reservation for the data service, and an aggregate data traffic load of $\rho_{\text{data}} = 6$ Erlang. Data calls of each priority class are equally likely to be generated. In the numerical experiment of Figure 4.12, the relative scheduling weight ϕ that differentiates between high- and low-priority data calls is varied between 0 and 1. Recall that the former setting implies *strict* priority differentiation and thus effectively turns the low-priority data class in a best effort class, while under the latter setting implies *no* priority differentiation and hence both data call types are treated identically. The figure displays the expected sojourn times for both data call classes versus the relative scheduling weight ϕ for a range of Weibullian data call size distributions with $\eta_{\text{data}} \in \{0, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16\}$.

For the high-priority data class, the trends induced by the range of data call size variabilities considered are in agreement with earlier observations: the greater the call size variability, the better the data QOS. For the low-priority data class, however, this qualitative trend fails to holds for very low values of the relative scheduling weight ϕ. Apparently, for $\phi \to 0$ and highly variable data call sizes, the time-varying process that governs the resource availability for low-priority data calls, becomes so unfavourable, with rather heavy tailed periods of (practically) no resources at all, that the data QOS degrades significantly. Observe further that as the relative scheduling weight ϕ increases from 0 to 1, the discrepancy between the high- and low-priority data QOS gradually disappears, with the expected sojourn times converging to those

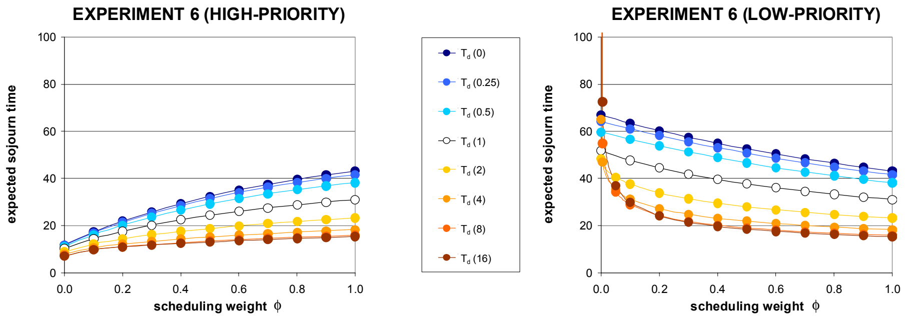


Figure 4.12 The expected high- (left) and low-priority (right) data call sojourn times versus the relative scheduling weight ϕ for a range of Weibullian data call size distributions.

found in Figures 4.3 (left; $\rho = 0.6$) and 4.11 (left). Moreover, apart from the rather extreme case of $\phi \approx 0$, the (relative) impact of the data call size variability increases with $\phi > 0$, most particularly for the high-priority data class.

4.8. CONCLUDING REMARKS

This chapter reports a sensitivity analysis of the impact of the data call size variability on the experienced data QOS in an integrated GSM/GPRS network, which is modelled as a Processor Sharing system with $d_{\max}$ service positions, with a time-varying capacity due to the arrival and termination of prioritised speech calls. The remarkable observation that the (conditional) expected sojourn time of a call is decreasing in the degree of variability of the data call size, is demonstrated by means of a series of simulation experiments. The QOS disparities between the different data call size distributions that were considered are most significant if the data traffic load ρ_{data} is high (given $d_{\max} = \infty$) and if the time scale at which the speech calls arrive and terminate is relatively large compared to the data traffic dynamics (large λ_{speech}, μ_{speech}). The disparities diminish as $\rho_{\text{data}} \to \infty$ (given $d_{\max} < \infty$) and as $\lambda_{\text{speech}}, \mu_{\text{speech}} \to \infty$ (keeping $\rho_{\text{speech}} \equiv \lambda_{\text{speech}}/\mu_{\text{speech}}$ fixed), with the (conditional) expected data call sojourn time converging to that experienced in an $M/G/1/PS$ queue whose fixed capacity is equal to the average remaining capacity in the integrated services model. Valuable insight into the validity of the presented observation is provided by means of an analytical treatment of extreme cases.

In view of the fact that the QOS in a FIFO system degrades under a greater data call size variability [213], while it is insensitive (fixed capacity) [203, 204, 213] or even improves (varying capacity) [this chapter] in a PS system, we have explicitly studied the trade-off between the FIFO and PS service disciplines in a extended system model. This model extends the basic model by queueing rather than rejecting data calls that find all service positions occupied upon arrival, in an infinite FIFO access queue. We have observed that in the extended model the impact of the data call size variability strongly depends on the number of service positions $d_{\max}^{\text{transfer}}$ in the transfer queue. In particular, for small $d_{\max}^{\text{transfer}}$ the FIFO access queue dominates and the QOS *degrades* under a greater data call size variability, while for large $d_{\max}^{\text{transfer}}$ the PS transfer queue dominates and the QOS *improves* under a greater data call size variability. We argue that in a well-dimensioned network the 'PS effect' typically dominates: the greater the data call size variability, the better the QOS.

A final numerical sensitivity analysis is motivated by the SHL model of Chapter 3 and concentrates on differentiation between two data priority classes in a PS model with varying capacity. This experiment demonstrates once again the phenomenon that a larger call size variability enhances the data QOS, *except* for the extreme case where low-priority data calls are treated in a(n approximately) 'best effort' manner, and the rather severe capacity variations cause the low-priority data QOS to degrade dramatically for a greater call size variability.

The principal relevance of the rather generic insights lies in the dimensioning and traffic management of integrated services telecommunications networks in general, e.g. aside from the considered context of GSM/GPRS networks, also for UMTS networks, next-generation WLANs, and fixed IP-based or ATM networks. In view of the commonly acknowledged property of e.g. WWW traffic to be heavy-tailed, the result indicates that assuming deterministic or lightly variable data call sizes, as is typically done for reasons of tractability in mathematical analyses or simulations, may lead to an underestimation of the experienced QOS. As a consequence, Call Admission Control schemes or network planning guidelines that are derived from such a model, are likely to be conservative.

CHAPTER 5

THROUGHPUT ANALYSIS OF PROCESSOR SHARING MODELS

PROCESSOR sharing queueing models are widely applicable to situations where a common resource is shared by a number of concurrent users. In particular, PS models have been fruitfully applied in the field of the performance evaluation of computer systems and telecommunication networks. For instance, the PS service discipline appropriately models the design principle of fair resource sharing by TCP controlled elastic data flows or packet level scheduling schemes in e.g. IP, GPRS, UMTS or Wireless Local Area Networks, as demonstrated throughout this monograph.

The 'classical' PS model consists of a single server fairly sharing its fixed capacity among the varying number of present calls. In light of emerging integrated services networks, a relevant extension that has been investigated in Chapters 2 and 3, is the PS queue with randomly varying service capacity, which models e.g. the impact of prioritised (speech) traffic on (low priority) video or data flows sharing a common network link. Important performance measures for PS queues are sojourn times and throughputs. In the queueing literature, the analyses of PS models are generally focussed towards the (conditional) expected *sojourn times* and their distribution, and many analytical results are available. Although the relevance is apparent from practical applications, *throughput* analyses are however rare and only a few results are known. The present chapter therefore concentrates on the analysis and comparison of a variety of relevant throughput measures in PS models with fixed or randomly varying capacity, motivated by the QOS analyses for integrated services second-generation cellular networks presented in Chapters 2 and 3. As was also the case for the sensitivity analysis of Chapter 4, the content of the current chapter is of a more theoretical nature, which is in line with the rather generic insights that are obtained.

The outline of the chapter is as follows. Section 5.1 discusses the scarce literature on throughput analysis in PS models, followed by a statement of contribution in Section 5.2. Section 5.3 specifies the characteristics of the considered services, the investigated PS models and the relevant performance measures. An analytical evaluation and comparison of the different throughput measures is presented in Section 5.4. Section 5.5 then discusses the results of an extensive set of numerical experiments carried out to provide further qualitative and quantitative insight into the throughput performance for the considered PS models. The concluding remarks in Section 5.6 end this chapter.

5.1. LITERATURE

In the literature the analysis of PS models is primarily targeted towards the expected *sojourn times* of data calls with a given service requirement (call size). Refer to Section 2.1 for an overview of the related literature.

Throughput analyses of PS models are extremely rare. Kherani and Kumar [127, 128] assess the PS service discipline as a model to evaluate the performance of TCP-controlled elastic (data) traffic in the Internet (see also [164, 175, 197]), and compare different throughput measures for the $M/G/1/PS$ model by means of analysis and simulations. In a multitude of reported investigations with a larger scope, e.g. dimensioning of large IP networks, a seemingly arbitrary throughput measure is selected as a basis for the performance analysis, without substantiating the validity of such a measure. A few references to studies where different throughput measures are applied, are pointed out when the different throughput measures are specified in Section 5.3.3.

5.2. CONTRIBUTION

The principal objective of this chapter is to derive and compare, both analytically and numerically, a variety of throughput performance measures in Processor Sharing models serving two distinct types of elastic calls. The principal merit of the 'throughput' as a QOS measure as that it rather generically applies to a range of elastic service types, and is in some sense normalised with respect to the size of e.g. data calls. Aside from *fixed* capacity PS systems, we also consider systems which integrate elastic traffic with prioritised stream (e.g. speech) traffic, thus establishing a system with *varying* service capacity from the elastic calls' perspective. Although the analysis is of a generic nature and the results are certainly more broadly applicable, the applied terminology and numerical experiments are associated with the example context of a single cell in an integrated services GSM/GPRS network.

Aside from a substantial original contribution in the definition, analysis and comparison of throughput measures, the few known results have been included in order to establish the survey character of the chapter. While from the customer's perspective, the *call-average throughput* is the most relevant throughput measure, in PS systems the call-average throughput may be hard to determine analytically, which is an important reason to assess the closeness of a number of other throughput measures. In several papers the *time-average throughput*, defined as the expected throughput the 'server' provides to an elastic call at an arbitrary (non-idle) time instant, or the *ratio* of the expected transfer volume and the expected sojourn time are applied to approximate the call-average throughput. In this chapter we introduce the *expected instantaneous throughput*, i.e. the throughput an admitted call experiences immediately upon admission to the system, as a new throughput measure, which can be analysed relatively easily. The experiments demonstrate that the newly proposed expected instantaneous throughput measure is the *only* measure which excellently approximates the call-average throughput for each of the investigated PS models and over the entire range of elastic traffic loads.

5.3. MODELS AND MEASURES

Consider a single GSM/GPRS cell with C traffic channels that are shared by speech and elastic (video or data) calls. The defining characteristics of the different services are similar to those specified in Chapter 3 and briefly summarised below, followed by the specification of the call handling procedures in four distinct system models. An overview of the considered performance measures ends the section.

5.3.1. TRAFFIC MODELS

Three distinct service types are considered in the investigated PS models:

SPEECH SERVICE: Speech calls arrive according to a Poisson process with arrival intensity λ_{speech}, have a generally distributed duration with mean $1/\mu_{\text{speech}}$, and require a fixed single-channel assignment. The speech traffic load is denoted $\rho_{\text{speech}} \equiv \lambda_{\text{speech}}/\mu_{\text{speech}}$.

VIDEO SERVICE: Video calls arrive according to a Poisson process with arrival intensity λ_{video}, have a generally distributed duration with mean $1/\mu_{\text{video}}$, and are elastic (*scalable*) in the ideal sense that the assigned number of traffic channels and thus the video quality can instantaneously and with perfect granularity adapt to the varying network load. The multislot restrictions of a GPRS terminal impose a maximum, denoted $\beta_{\text{GPRS}}^{\max}$, on the number of traffic channels that can be assigned to a video call. On the other hand, acceptable video quality is guaranteed by means of a minimum channel assignment of $\beta_{\text{video}}^{\min} \in [0, \beta_{\text{GPRS}}^{\max}]$ traffic channels, corresponding to a bit rate of $r_{\text{video}}\beta_{\text{video}}^{\min}$ kbits/s, with r_{video} the effective video bit rate per traffic channel. The video traffic load is defined as $\rho_{\text{video}} \equiv \lambda_{\text{video}}/\mu_{\text{video}}$.

DATA SERVICE: Data calls arrive according to a Poisson process with arrival intensity λ_{data}. A data call is assumed to be the transfer of a file with a generally distributed size, which is expressed in its nominal sojourn time assuming a single dedicated traffic channel, by means of a normalisation with respect to the effective data bit rate per traffic channel r_{data} (in kbits/s). The mean file size of $1/\mu_{\text{data}}$ thus corresponds with an actual transfer volume of $r_{\text{data}}/\mu_{\text{data}}$ kbits. The delay-tolerant data calls are also elastic in that they can tolerate a varying channel assignment, which affects the experienced throughput and thus the data call's sojourn time. The (potential) minimum guaranteed data rate and the multislot restrictions limit the data calls' channel assignments to the range $\left[\beta_{\text{data}}^{\min}, \beta_{\text{GPRS}}^{\max}\right]$. The (normalised) data traffic load is given by $\rho_{\text{data}} \equiv \lambda_{\text{data}}/\mu_{\text{data}}$ ($\rho_{\text{data}}^{\star} \equiv \rho_{\text{data}}/C$).

5.3.2. SYSTEM MODELS

Four distinct performance models are investigated, concentrating on one of the specific elastic services, to be handled according to a Processor Sharing service discipline with a fixed or varying capacity. In the latter case, the considered elastic service shares the aggregate capacity with a speech service, which utilises the cell's capacity with preemptive priority, and thus implicitly leaves a time-varying residual capacity for the elastic calls. The different models, denoted V, D, SV and SD, are special instances of the more general SVD model investigated in Chapter 3. The focussed models allow a more comprehensive analysis, however, and are furthermore selected to convey the principal results most transparently. A brief specification of the models is given below. Let $S(t)$, $V(t)$ and $D(t)$ denote the process following the number of speech, video and data calls present at time $t \geq 0$, with states denoted s, v and d, respectively.

V MODEL: In the V model video calls share the available C channels (fixed) in a PS fashion, i.e. given a presence of v video calls, each video call enjoys an instantaneous channel assignment of $\beta_{\text{video}}(v) \equiv \min\{C/v,\ \beta_{\text{GPRS}}^{\max}\}$, obeying the multislot restriction enforced by $\beta_{\text{GPRS}}^{\max}$. In case of a positive minimum QOS requirement $\beta_{\text{video}}^{\min} > 0$, Call Admission Control enforces a maximum presence of $v_{\max} \equiv \left\lfloor C/\beta_{\text{video}}^{\min} \right\rfloor$ video calls.

SV MODEL: In the SV model the C traffic channels are dynamically shared by speech and video calls. Aside from the channels that are assigned to present video calls in order to meet their QOS requirement, the capacity is available with preemptive priority for speech calls. In other words, an arriving speech call is admitted if and only if $s+1 \leq s_{\max}(v) \equiv \left\lfloor C - v\beta_{\text{video}}^{\min} \right\rfloor$, given a presence of s speech and v video calls. Analogously, if $\beta_{\text{video}}^{\min} > 0$, the condition for the admission of a video call is given by $v+1 \leq v_{\max}(s) \equiv \left\lfloor (C-s)/\beta_{\text{video}}^{\min} \right\rfloor$. At any given time, the capacity that is not assigned to speech calls, is fairly shared by the present video calls in a PS fashion, i.e. each video call is assigned an instantaneous channel assignment of $\beta_{\text{video}}(s,v) \equiv \min\{(C-s)/v, \beta_{\text{GPRS}}^{\max}\}$, which is guaranteed to exceed the minimum QOS requirement due to effects of the Call Admission Control. Observe that the SV model is an example of a multi-rate model (see e.g. [124, 194]) incorporating speech and video calls with respective capacity requirements of 1 and $\beta_{\text{video}}^{\min}$ traffic channels.

D MODEL: The D model is equivalent to the $M/G/1/d_{\max}/GPS$ queueing model with state-dependent aggregate service rates (due to $\beta_{\text{GPRS}}^{\max}$) treated in [54], i.e. given a presence of d data calls, each such data call is assigned an instantaneous channel assignment of $\beta_{\text{data}}(d) \equiv \min\{C/d, \beta_{\text{GPRS}}^{\max}\}$. The Call Admission Control threshold $d_{\max} \equiv \left\lfloor C/\beta_{\text{data}}^{\min} \right\rfloor$ is enforced if $\beta_{\text{data}}^{\min} > 0$.

SD MODEL: In the SD model the C traffic channels are dynamically shared by speech and data calls in a similar manner as in the SV model. The Call Admission Control conditions for the admission of a speech or data call are given by $s+1 \leq s_{\max}(d) \equiv \left\lfloor C - d\beta_{\text{data}}^{\min} \right\rfloor$ and $d+1 \leq \left\lfloor (C-s)/\beta_{\text{data}}^{\min} \right\rfloor$ (only if $\beta_{\text{data}}^{\min} > 0$), respectively, given a presence of s speech and d data calls. At any given time, the capacity that is not assigned to speech calls, is fairly shared by the present data calls, i.e. each data call is assigned an instantaneous channel assignment of $\beta_{\text{data}}(s,d) \equiv \min\{(C-s)/d, \beta_{\text{GPRS}}^{\max}\} \geq \beta_{\text{data}}^{\min}$.

Note that no access queue is maintained in the (S)D model to temporarily store data calls that cannot be assigned resources immediately upon arrival (only if $\beta_{\text{data}}^{\min} > 0$). Rather, such calls are blocked and cleared from the system.

5.3.3. PERFORMANCE MEASURES

In this subsection the definitions of the different performance measures are given. The definitions are formulated in a generic manner and thus apply to both video and data calls. Denote with a_k (d_k) the arrival (departure) time of the k^{th} *admitted* elastic call, with $\tau_k \equiv d_k - a_k$ the call's sojourn time and with x_k the associated information volume (in kbits) transferred during its sojourn. Recall that for the video service the durations τ_k are autonomously sampled and the transfer volumes x_k are determined by the system dynamics, while for the data service the reverse holds. Let τ and x be the corresponding generic random variables with expected values $\mathbf{E}\{\tau\}$ and $\mathbf{E}\{x\}$. The *call-average* throughput is defined as

$$\mathbf{R}^c \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{x_k}{\tau_k} = \mathbf{E}\left\{\frac{x}{\tau}\right\}. \tag{5.1}$$

With $N(t)$ the number of elastic calls present in the system and $C(t)$ the aggregate number of channels assigned to the elastic service at time $t \geq 0$, the *time-average* throughput is defined as

$$\mathbf{R}^t \equiv \lim_{t\to\infty} \frac{\frac{1}{t}\int\limits_0^t \frac{rC(u)}{N(u)} 1\{N(u) \geq 1\}\, du}{\frac{1}{t}\int\limits_0^t 1\{N(u) \geq 1\}\, du}, \tag{5.2}$$

where r denotes the effective (service-specific) information bit rate per traffic channel. Note that $N(t)$ is given by $V(t)$ in the (S)V model or $D(t)$ in the (S)D model, while $C(t)/N(t)$ is given by the channel assignment functions $\beta(\cdot)$. The time-average throughput is used to approximate the call-average throughput in e.g. [64, 127, 128]. We introduce the expected *instantaneous* throughput as

$$\mathbf{R}^i \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{rC(a_k)}{N\left(a_k^+\right)}, \tag{5.3}$$

where $N\left(a_k^+\right)$ denotes the number of present elastic calls immediately after the k^{th} elastic call arrival and thus includes the new call. The *ratio* of the expected transfer volume and the expected sojourn time is defined as

$$\mathbf{R}^r \equiv \lim_{n\to\infty} \frac{\frac{1}{n}\sum\limits_{k=1}^{n} x_k}{\frac{1}{n}\sum\limits_{k=1}^{n} \tau_k} = \frac{\mathbf{E}\left\{x\right\}}{\mathbf{E}\left\{\tau\right\}}, \tag{5.4}$$

which is applied in e.g. [16, 17, 30, 31, 61, 190]. Note that $\mathbf{R}^r$ can also be written as

$$\mathbf{R}^r = \frac{\lambda\left(1-\mathbf{P}\right)\mathbf{E}\left\{x\right\}}{\lambda\left(1-\mathbf{P}\right)\mathbf{E}\left\{\tau\right\}} = \lim_{t\to\infty} \frac{\frac{1}{t}\int\limits_0^t rC\left(u\right)du}{\frac{1}{t}\int\limits_0^t N\left(u\right)du},$$

where λ denotes the (generic) elastic call arrival rate and $\mathbf{P}$ the elastic call blocking probability (see also below). This alternate expression for $\mathbf{R}^r$ is given by the ratio of the long-term average aggregate system throughput and the long-term average number of elastic calls in the system. Its equivalence to expression (5.4) is due to the fact that in equilibrium the aggregate admitted bit rate must be equal to the aggregate processed bit rate (numerator) and Little's law (denominator). As a final measure, the (unitless) call-average *stretch* (or the *normalised sojourn time*) is given by

$$\mathbf{S} \equiv \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \frac{\tau_k}{\left(\frac{x_k}{rC}\right)} = rC\,\mathbf{E}\left\{\frac{\tau}{x}\right\}, \tag{5.5}$$

which is relevant for the data service only and is used as a performance measure in e.g. [121, 197]. For the special case of unrestricted channel assignments, i.e. $\beta_{\text{data}}^{\text{min}} = \beta_{\text{video}}^{\text{min}} = 0$ and $\beta_{\text{GPRS}}^{\text{max}} \geq C$, let $\widetilde{\mathbf{R}}^c$, $\widetilde{\mathbf{R}}^t$, $\widetilde{\mathbf{R}}^i$, $\widetilde{\mathbf{R}}^r$ and $\widetilde{\mathbf{S}}$ denote the associated performance measures corresponding to the more general measures specified above.

In addition to these throughput measures, the included Call Admission Control schemes imply the occurrence of call blocking and thus the need to determine the speech, video and data call blocking probability (denoted $\mathbf{P}$), defined as the probability that an arriving call of a given type is denied admission to the system. Clearly, video or data calls experience blocking only if $\beta_{\text{video}}^{\text{min}} > 0$ or $\beta_{\text{data}}^{\text{min}} > 0$, respectively.

5.4. PERFORMANCE ANALYSIS

In this section we derive analytical expressions for the relevant performance measures in the four models specified above.

5.4.1. SV MODEL

Consider the SV model with generally distributed speech and video call durations. The evolution of the system in the SV model can then be described by the continuous-time stochastic process $(S(t), V(t))_{t\geq 0}$, with states denoted (s, v). The process' state space is given by $\mathbb{S} \equiv \left\{(s,v) \in \mathbb{N}_0 \times \mathbb{N}_0 : s + v\beta_{\text{video}}^{\text{min}} \leq C\right\}$. The unique equilibrium probability vector $\boldsymbol{\pi}$ of the stochastic process, given by

$$\pi\left(s,v\right) = \left(\sum_{(s,v)\in\mathbb{S}} \frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\right)^{-1} \frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!},\ \ (s,v)\in\mathbb{S},$$

is *insensitive* to the specific form of the speech and video call distributions, depending on their means only (see e.g. [124, 125, 194]). For the special case of unrestricted channel assignments to the video service, the state space is equal to $\widetilde{\mathbb{S}} \equiv \{(s,v) \in \mathbb{N}_0 \times \mathbb{N}_0 : s \leq C\}$, and the equilibrium distribution is given by the product form

$$\widetilde{\pi}\left(s,v\right) = \exp\left(-\rho_{\text{video}}\right)\left(\sum_{s=0}^{C} \frac{\rho_{\text{speech}}^{s}}{s!}\right)^{-1} \frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!},\ \ (s,v)\in\mathbb{S}.$$

Using the PASTA property [224], the call blocking probabilities are readily derived from the equilibrium distribution:

$$\mathbf{P}_{\text{speech}} = \sum_{v=0}^{v_{\max}(0)} \pi\left(s_{\max}\left(v\right),v\right) \text{ and } \mathbf{P}_{\text{video}} = \sum_{s=0}^{C} \pi\left(s, v_{\max}\left(s\right)\right).$$

In the case of unrestricted channel assignments, the speech call blocking probability is simply given by the Erlang loss probability, since speech traffic does not 'see' video traffic in the absence of video QOS guarantees, while the video call blocking probability equals zero.

CALL-AVERAGE THROUGHPUT

We begin the video throughput analysis with a conditional analysis of the *call-average* throughput of a video call of a given duration τ which is admitted to the system in state (s, v). In this conditional analysis we confine ourselves to the case of *exponentially* distributed speech and video call durations and recall the conditional video QOS analysis of the SVD model in Section 3.6.1.

For each state $(s, v) \in \mathbb{S}^{+}_{\text{video}} \equiv \{(s, v) \in \mathbb{S} : v > 0\}$, denote with $\widehat{x}_{s,v}(\tau)$ the conditional expected transfer volume of an admitted video call of duration τ, arriving at a given system state (s, v), where v includes the new video call. The derivation involves a modified version of the Markov chain that is readily specified to describe the evolution of the SV model's stochastic process under the exponentiality assumption (see also Chapter 3). Characterised by the presence of one permanent video call, the modified Markov chain consequently has the reduced state space $\mathbb{S}^{+}_{\text{video}}$. The video call departure rates in the associated infinitesimal generator $\mathcal{Q}^{\star}_{\text{video}}$ reflect the presence of the permanent video call, i.e. $\mathcal{Q}^{\star}_{\text{video}}\left((s, v)\,;(s, v-1)\right) = (v-1)\mu_{\text{video}}$. The equilibrium distribution vector $\boldsymbol{\pi}^{\star}_{\text{video}} \equiv \left(\pi^{\star}_{\text{video}}(s, v),\ (s, v) \in \mathbb{S}^{+}_{\text{video}}\right)$ of the modified Markov chain is, invoking reversibility and truncation of a reversible process [125], readily obtained as

$$\pi^{\star}_{\text{video}}\left(s, v\right) = \frac{\pi\left(s, v-1\right)}{\sum\limits_{(s',v')\in\mathbb{S}^{+}_{\text{video}}} \pi\left(s', v'-1\right)},\ \ (s, v) \in \mathbb{S}^{+}_{\text{video}}, \tag{5.6}$$

i.e. the equilibrium probabilities $\pi^{\star}_{\text{video}}(s, v)$ corresponding to the modified Markov chain with one permanent video call are *equal* to the conditional probabilities that a newly admitted video call brings the system in state (s, v) in the original Markov chain. The equilibrium distribution $\boldsymbol{\pi}^{\star}_{\text{video}}$ is also known to be insensitive to the specific form of the speech and video call distributions [124, 125, 194]. Let $\mathcal{B}_{\text{video}} \equiv$ *diag*$(\beta_{\text{video}}(s, v),\ (s, v) \in \mathbb{S}^{+}_{\text{video}})$ denote the diagonal matrix of video channel assignments, lexicographically ordered in (s, v).

As a special case of Proposition 3.1, for exponentially distributed video call durations the conditional expected video transfer volume vector $\widehat{\mathbf{x}}(\tau) \equiv (\widehat{x}_{s,v}(\tau),\ (s, v) \in \mathbb{S}^{+}_{\text{video}})$ is then given by

$$\widehat{\mathbf{x}}(\tau) = \tau r_{\text{video}}\left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}},$$

where $\boldsymbol{\gamma}_{\text{video}} \equiv \left(\gamma_{\text{video}}(s,v),\ (s,v) \in \mathbb{S}^+_{\text{video}}\right)$ is the unique solution to

$$\begin{aligned} \mathcal{Q}^\star_{\text{video}} \boldsymbol{\gamma}_{\text{video}} &= r_{\text{video}} \left\{ \left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1} \right) \mathbf{1} - \mathcal{B}_{\text{video}} \mathbf{1} \right\}, \\ \boldsymbol{\pi}^\star \boldsymbol{\gamma}_{\text{video}} &= \mathbf{0}. \end{aligned} \tag{5.7}$$

The conditional expected (call-average) video throughput $\mathbf{R}^c_{\text{video}}(s,v,\tau)$ of a video call admitted to the system in state (s,v) and with a given holding time τ is then given by

$$\mathbf{R}^c_{\text{video}}(s,v,\tau) = \frac{\widehat{x}_{s,v}(\tau)}{\tau} \tag{5.8}$$

(recall (5.1)), while deconditioning on the system state upon admission yields the conditional expected (call-average) video throughput of an admitted video call with duration τ, given by

$$\begin{aligned} \mathbf{R}^c_{\text{video}}(\tau) &= \sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s,v-1)}{\sum\limits_{(s',v') \in \mathbb{S}^+_{\text{video}}} \pi(s',v'-1)} \right) \mathbf{R}^c_{\text{video}}(s,v,\tau) \\ &= \boldsymbol{\pi}^\star_{\text{video}} \left\{ r_{\text{video}} \left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1} \right) \mathbf{1} + \frac{1}{\tau} \left[\mathcal{I} - \exp\left\{ \tau \mathcal{Q}^\star_{\text{video}} \right\} \right] \boldsymbol{\gamma}_{\text{video}} \right\} \\ &= r_{\text{video}} \left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1} \right) + \frac{1}{\tau} \boldsymbol{\pi}^\star_{\text{video}} \left(\boldsymbol{\gamma}_{\text{video}} - \sum_{k=0}^{\infty} \frac{\left(\tau \mathcal{Q}^\star_{\text{video}} \right)^k}{k!} \boldsymbol{\gamma}_{\text{video}} \right) \\ &= r_{\text{video}} \boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1} \\ &= r_{\text{video}} \sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s,v-1)}{\sum\limits_{(s',v') \in \mathbb{S}^+_{\text{video}}} \pi(s',v'-1)} \right) \beta_{\text{video}}(s,v), \end{aligned}$$

using (5.7) and $\boldsymbol{\pi}^\star_{\text{video}} \mathcal{Q}^\star_{\text{video}} = \mathbf{0}$. Observe that $r_{\text{video}} \boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}$ is equal to the time-average video throughput in the SV model with one permanent video call (see also below). Comparing the first and last expression in the above derivation might confuse the reader into thinking that $\mathbf{R}^c_{\text{video}}(s,v,\tau)$ is simply equal to $r_{\text{video}} \beta_{\text{video}}(s,v)$, which is however readily seen to be not the case. Observe that $\mathbf{R}^c_{\text{video}}(\tau)$ does not depend on τ, so that the call-average video throughput is given by

$$\mathbf{R}^c_{\text{video}} = \int\limits_{\tau=0}^{\infty} \mathbf{R}^c_{\text{video}}(\tau)\mu_{\text{video}} \exp\left\{-\tau\mu_{\text{video}}\right\} d\tau = \mathbf{R}^c_{\text{video}}(\tau) = r_{\text{video}}\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}. \tag{5.9}$$

Remark 5.1 In the conditional video throughput analysis of the SVD model in Chapter 3 the above deconditioning step does *not* work out so nicely, due the fact that in the SVD model the equilibrium probabilities $\pi^{\star}_{\text{video}}(s, v, d)$ corresponding to the modified Markov chain with one permanent video call are *not* equal to the conditional probabilities that a newly admitted video call brings the system in state (s, v, d) in the original Markov chain, unlike in the reversible SV model considered here.

Whereas the above derivations utilised the exponentiality of the speech and video call durations, Proposition 5.1 claims that the obtained expressions for both $\mathbf{R}^c_{\text{video}}$ and $\mathbf{R}^c_{\text{video}}(\tau)$ (*not* $\mathbf{R}^c_{\text{video}}(s, v, \tau)$) are *insensitive* to the distributions of the speech and video call durations, apart from their means.

Proposition 5.1 *The call-average video throughput* $\mathbf{R}^c_{video}$ *and the conditional call-average video throughput* $\mathbf{R}^c_{video}(\tau)$ *are insensitive to the speech and video call duration distributions apart from their means.*

Proof The stationary joint distribution $\pi\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right)$ of the number of speech (S) and video calls (V) present in the system and the associated residual call durations $\boldsymbol{\Theta}_{\text{speech}} \equiv \left(\Theta_{\text{speech}}(1), \cdots, \Theta_{\text{speech}}(S)\right)$ and $\boldsymbol{\Theta}_{\text{video}} \equiv \left(\Theta_{\text{video}}(1), \cdots, \Theta_{\text{video}}(V)\right)$ is given by (see e.g. [62])

$$\begin{aligned}
&\pi\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right) \\
&\quad \equiv \Pr\left\{S = s, V = v, \boldsymbol{\Theta}_{\text{speech}} \in \left[\boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{speech}} + d\boldsymbol{\vartheta}_{\text{speech}}\right],\right. \\
&\qquad\qquad \left.\boldsymbol{\Theta}_{\text{video}} \in \left[\boldsymbol{\vartheta}_{\text{video}}, \boldsymbol{\vartheta}_{\text{video}} + d\boldsymbol{\vartheta}_{\text{video}}\right]\right\} \\
&\quad = G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C\right)\left\{\frac{\rho^s_{\text{speech}}}{s!}\frac{\rho^v_{\text{video}}}{v!}\prod_{s'=1}^{s}\left(\frac{\overline{\Phi}_{\text{speech}}\left(\vartheta_{\text{speech}}\left(s'\right)\right)}{\mu^{-1}_{\text{speech}}}d\vartheta_{\text{speech}}\left(s'\right)\right)\times\right. \\
&\qquad\qquad \left.\prod_{v'=1}^{v}\left(\frac{\overline{\Phi}_{\text{video}}\left(\vartheta_{\text{video}}\left(v'\right)\right)}{\mu^{-1}_{\text{video}}}d\vartheta_{\text{video}}\left(v'\right)\right)\right\},
\end{aligned}$$

for $(s,v) \in \mathbb{S} = \mathbb{S}(C) \equiv \left\{(s,v) \in \mathbb{N}_0 \times \mathbb{N}_0 : s + v\beta_{\text{video}}^{\min} \leq C\right\}$, $\boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}} \geq \mathbf{0}$, where the vectors $d\boldsymbol{\vartheta}_{\text{speech}}$ and $d\boldsymbol{\vartheta}_{\text{video}}$ consist of infinitesimally small elements,

$$G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C\right) \equiv \left(\sum_{(s,v)\in\mathbb{S}(C)} \frac{\rho_{\text{speech}}^{s}}{s!} \frac{\rho_{\text{video}}^{v}}{v!}\right)^{-1},$$

and where $\overline{\Phi}_{\text{speech}}$ and $\overline{\Phi}_{\text{video}}$ denote the complementary cumulative distributions of the speech and video call durations, respectively.

Using PASTA, the joint distribution $\pi_{\text{video}}^{\bullet}(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}})$ of $(S, V, \boldsymbol{\Theta}_{\text{speech}}, \boldsymbol{\Theta}_{\text{video}})$ upon *admission* of a tagged video call is readily given by

$$\begin{aligned}
&\pi_{\text{video}}^{\bullet}(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}) \\
&\quad \equiv \Pr\{S = s, V = v, \boldsymbol{\Theta}_{\text{speech}} \in [\boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{speech}} + d\boldsymbol{\vartheta}_{\text{speech}}], \\
&\qquad\quad \boldsymbol{\Theta}_{\text{video}} \in [\boldsymbol{\vartheta}_{\text{video}}, \boldsymbol{\vartheta}_{\text{video}} + d\boldsymbol{\vartheta}_{\text{video}}] \mid s + v\beta_{\text{video}}^{\min} \leq C - \beta_{\text{video}}^{\min}\Big\} \\
&\quad = G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C - \beta_{\text{video}}^{\min}\right) \times \\
&\qquad\quad \left\{\frac{\rho_{\text{speech}}^{s}}{s!} \frac{\rho_{\text{video}}^{v}}{v!} \prod_{s'=1}^{s} \frac{\overline{\Phi}_{\text{speech}}(\vartheta_{\text{speech}}(s'))}{\mu_{\text{speech}}^{-1}} \prod_{v'=1}^{v} \frac{\overline{\Phi}_{\text{video}}(\vartheta_{\text{video}}(v'))}{\mu_{\text{video}}^{-1}}\right\},
\end{aligned}$$

for $(s,v) \in \mathbb{S}\left(C - \beta_{\text{video}}^{\min}\right)$, where v *excludes* the newly admitted tagged video call.

Observe that $\pi_{\text{video}}^{\bullet}(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}})$ is equal to the stationary joint distribution of the number of speech and video calls and their residual call durations in a corresponding system with capacity $C - \beta_{\text{video}}^{\min}$ instead of C, or equivalently, in the original system but with one *permanent* video call (where v *excludes* this call). Hence the system state remains stochastically identical throughout the duration of the tagged video call. The associated (partially deconditioned) system state distribution $\pi_{\text{video}}^{\bullet}(s,v)$ is given by

$$\pi_{\text{video}}^{\bullet}(s,v) = \int_{\vartheta_{\text{speech}}(1)=0}^{\infty} \cdots \int_{\vartheta_{\text{video}}(v)=0}^{\infty} \pi_{\text{video}}^{\bullet}(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}})$$

$$= G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C - \beta_{\text{video}}^{\min}\right)\left\{\frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\right\}, \tag{5.10}$$

for $(s, v) \in \mathbb{S}\left(C - \beta_{\text{video}}^{\min}\right)$. Since the throughput of the tagged video call is completely determined by the distribution of the number of speech and *other* video calls present during its lifetime, as given in (5.10), it is then immediately clear that the conditional call-average throughput $\mathbf{R}_{\text{video}}^{c}(\tau)$ of the tagged video call is *independent* of its duration τ, i.e. $\mathbf{R}_{\text{video}}^{c}(\tau) = \mathbf{R}_{\text{video}}^{c}$, for all $\tau \geq 0$. In particular, it is equal to the expected instantaneous video throughput experienced upon admission, which inherits its insensitivity from the insensitivity of $\pi_{\text{video}}^{\star}$ (see also Section 5.4.1 below). □

Remark 5.2 The stationary probability $\pi_{\text{video}}^{\bullet}(s, v)$ given in expression (5.10), where v *excludes* the tagged (permanent) video call, is readily verified to be equivalent to the conditional probability $\pi_{\text{video}}^{\star}(s, v+1)$ given in expression (5.6), where v *includes* the newly admitted video call.

For the case without channel assignment restrictions it is readily derived that

$$\widetilde{\mathbf{R}}_{\text{video}}^{c} = r_{\text{video}}\frac{1 - \exp\left(-\rho_{\text{video}}\right)}{\rho_{\text{video}}}\left(C - \rho_{\text{speech}}\left(1 - \mathbf{P}_{\text{speech}}\right)\right).$$

TIME-AVERAGE THROUGHPUT

Using the theory of regenerative processes (e.g. [213, 224]), the *time-average* video throughput is given by

$$\begin{aligned}\mathbf{R}_{\text{video}}^{t} &= \lim_{t \to \infty} \frac{\frac{1}{t}\int_{0}^{t} r_{\text{video}}\beta_{\text{video}}\left(S(u), V(u)\right)1\left\{V(u) \geq 1\right\}du}{\frac{1}{t}\int_{0}^{t} 1\left\{V(u) \geq 1\right\}du} \\ &= r_{\text{video}}\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}}\left(\frac{\pi(s,v)}{\sum\limits_{(s',v')\in\mathbb{S}_{\text{video}}^{+}}\pi(s',v')}\right)\beta_{\text{video}}(s,v),\end{aligned} \tag{5.11}$$

(cf. (5.2)), where $\pi(s,v)/\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}}\pi(s,v)$ is the equilibrium probability that the system is in state (s, v), conditioned on the presence of at least one video call. The involved Césaro limits are derived using the renewal reward theorem [213, 224]. For the special case without channel assignment restrictions this yields

$$\widetilde{\mathbf{R}}^{t}_{\text{video}} = \frac{r_{\text{video}}}{\left(\exp\left(\rho_{\text{video}}\right) - 1\right)} \left(\sum_{v=1}^{\infty} \frac{\rho^{v}_{\text{video}}}{vv!} \right) \left(C - \rho_{\text{speech}}\left(1 - \mathbf{P}_{\text{speech}}\right)\right),$$

where $\mathbf{P}_{\text{speech}}$ is the Erlang loss probability. Note that the derivation of (5.11) does not require information on the specific form of the equilibrium distribution π. As this equilibrium distribution is insensitive to the call duration distribution (except for its mean), this property is inherited by the time-average video throughput.

EXPECTED INSTANTANEOUS THROUGHPUT

The expected *instantaneous* video throughput as defined in (5.3) is obtained as

$$\begin{aligned}
\mathbf{R}^{i}_{\text{video}} &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} r_{\text{video}} \beta_{\text{video}}\left(S\left(a_k\right), V\left(a_k^{+}\right)\right) \\
&= r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \left(\frac{\pi(s, v-1)}{\sum_{(s',v')\in\mathbb{S}^{+}_{\text{video}}} \pi(s', v'-1)} \right) \beta_{\text{video}}\left(s, v\right), \\
&= r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \pi^{\star}_{\text{video}}\left(s, v\right) \beta_{\text{video}}\left(s, v\right), \qquad (5.12)
\end{aligned}$$

once again applying the theory of regenerative processes. As for the time-average throughput, the expected instantaneous video throughput measure inherits its insensitivity with respect to the specific form of the video call duration distribution from the insensitivity of $\pi^{\star}_{\text{video}}$. Observe that the expected instantaneous video throughput is equal to the call-average video throughput, and hence so is the special case with unrestricted channel assignments.

RATIO THROUGHPUT MEASURE

The *ratio* of the expected video call transfer volume and the expected video call duration is given by

$$\mathbf{R}^r_{\text{video}} = \frac{\mathbf{E}\left\{\tau \mathbf{R}^c_{\text{video}}(\tau)\right\}}{\mu^{-1}_{\text{video}}} = \mathbf{R}^c_{\text{video}}$$

(cf. (5.4)), where the numerator is indeed equal to the expected transfer volume of a video call, using the fact that $\mathbf{R}^c_{\text{video}}(\tau) = \mathbf{R}^c_{\text{video}}$ does not depend on τ. It is readily seen that also for the special case of unrestricted channel assignments, the ratio throughput measure is equal to the corresponding call-average video throughput.

COMPARISON OF THROUGHPUT MEASURES

The *call-average* video throughput, the expected *instantaneous* video throughput and the *ratio* of the expected video call transfer volume and the expected video call duration appear to be identical, i.e.

$$\mathbf{R}^c_{\text{video}} = \mathbf{R}^i_{\text{video}} = \mathbf{R}^r_{\text{video}},$$

and hence what remains is to compare these measures with the *time-average* throughput. Based on the explicit expressions (5.9) and (5.11), we will show in Proposition 5.2 for the case of $\beta^{\min}_{\text{video}} \in \{0, 1, \cdots, C\}$ that the time-average throughput exceeds the call-average throughput: $\mathbf{R}^c_{\text{video}} \le \mathbf{R}^t_{\text{video}}$, noting that numerical evaluations indicate that the inequality also holds for non-integer $\beta^{\min}_{\text{video}}$, although the presented line of proof does not work for this more general case. As an interesting corollary, we obtain that the time-average video throughput is monotonous in the offered video traffic load, i.e. $\frac{\partial \mathbf{R}^t_{\text{video}}}{\partial \rho_{\text{video}}} \le 0$, which is noted to be non-trivial. While for $\rho_{\text{speech}} = 0$ (V model) this monotonicity can readily be concluded via stochastic monotonicity, for $\rho_{\text{speech}} > 0$ speech calls may take the place of video calls thus destroying stochastic monotonicity.

Proposition 5.2 *In the* SV *model with $\beta^{\min}_{\text{video}} \in \{0, 1, \cdots, C\}$, the call-average video throughput is less than or equal to the time-average video throughput: $\mathbf{R}^c_{video} \le \mathbf{R}^t_{video}$.*

Proof For the extreme cases of infinitesimally small or infinitely large video traffic loads, it is readily argued that the call- and time-average video throughput measures are identical. Under an extremely *light* video traffic load ($\rho_{\text{video}} \downarrow 0$), a (rarely) occurring system state $(s, v) \in \mathbf{S}^+_{\text{video}}$ must have $v = 1$, almost surely, for both the original stochastic process, and the modified process with one permanent video call. As a consequence, the time-average video throughputs of both processes are identical, and hence so are the call- and time-average video throughputs of the original process. We thus have that

$$\lim_{\rho_{\text{video}} \downarrow 0} \mathbf{R}^t_{\text{video}} = \lim_{\rho_{\text{video}} \downarrow 0} \mathbf{R}^c_{\text{video}}$$

as can readily be verified from (5.9) and (5.11).

Alternatively, an infinitely *heavy* video traffic load ($\rho_{\text{video}} \to \infty$, assuming $\beta^{\min}_{\text{video}} > 0$ for stability) leads to a (complete or near) crowding out of speech calls, and implies the permanent presence of $v_{\max}(0) = \left\lfloor C_{\text{total}}/\beta^{\min}_{\text{video}} \right\rfloor \geq 1$ video calls, and hence again the performance of the original and the modified process are the same. In particular, all video throughput measures are identical, so that

$$\lim_{\rho_{\text{video}} \to \infty} \mathbf{R}^t_{\text{video}} = \lim_{\rho_{\text{video}} \to \infty} \mathbf{R}^c_{\text{video}}.$$

Now assume that $0 < \rho_{\text{video}} < \infty$. Then, from (5.9) and (5.11),

$$\begin{aligned}
&\mathbf{R}^c_{\text{video}} \leq \mathbf{R}^t_{\text{video}} \\
&\quad \Longleftrightarrow r_{\text{video}} \sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s, v-1)}{\sum\limits_{(s',v') \in \mathbb{S}^+_{\text{video}}} \pi(s', v'-1)} \right) \beta_{\text{video}}(s, v) \\
&\qquad\quad \leq r_{\text{video}} \sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s, v)}{\sum\limits_{(s',v') \in \mathbb{S}^+_{\text{video}}} \pi(s', v')} \right) \beta_{\text{video}}(s, v) \\
&\quad \Longleftrightarrow \left(\sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^{v-1}_{\text{video}}}{s!\,(v-1)!} \beta_{\text{video}}(s, v) \right) \left(\sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^v_{\text{video}}}{s!v!} \right) + \\
&\qquad\quad - \left(\sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^v_{\text{video}}}{s!v!} \beta_{\text{video}}(s, v) \right) \left(\sum_{(s,v) \in \mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^{v-1}_{\text{video}}}{s!\,(v-1)!} \right) \leq 0
\end{aligned}$$

$$\begin{aligned}
\iff & \left(\sum_{v=0}^{v_{\max}-1}\rho_{\text{video}}^{v}\sum_{s=0}^{C-\beta^{\min}(v+1)}\frac{\rho_{\text{speech}}^{s}}{s!v!}\beta_{\text{video}}\left(s,v+1\right)\right)\times \\
& \left(\sum_{w=0}^{v_{\max}-1}\rho_{\text{video}}^{w}\sum_{s=0}^{C-\beta^{\min}(w+1)}\frac{\rho_{\text{speech}}^{s}}{s!\left(w+1\right)!}\right)+ \\
& -\left(\sum_{v=0}^{v_{\max}-1}\rho_{\text{video}}^{v}\sum_{s=0}^{C-\beta^{\min}(v+1)}\frac{\rho_{\text{speech}}^{s}}{s!\left(v+1\right)!}\beta_{\text{video}}\left(s,v+1\right)\right)\times \\
& \left(\sum_{w=0}^{v_{\max}-1}\rho_{\text{video}}^{w}\sum_{s=0}^{C-\beta^{\min}(w+1)}\frac{\rho_{\text{speech}}^{s}}{s!w!}\right)\leq 0
\end{aligned}$$

Recognising that the left-hand side is a polynomial in ρ_{video} of degree $2\left(v_{\max}-1\right)=2\left(\left\lfloor\frac{C}{\beta_{\text{video}}^{\min}}\right\rfloor-1\right)$, the above condition can be written in the following form:

$$\sum_{k=0}^{2(v_{\max}-1)}\rho_{\text{video}}^{k}\sum_{v+w=k}\zeta_{v,w}\leq 0, \tag{5.13}$$

where the coefficients $\zeta_{v,w}$, $v,w,=0,\cdots,v_{\max}-1$, are given by

$$\begin{aligned}
\zeta_{v,w} &= \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)}\frac{\rho_{\text{speech}}^{s}}{s!v!}\beta_{\text{video}}\left(s,v+1\right)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)}\frac{\rho_{\text{speech}}^{s}}{s!\left(w+1\right)!}\right)+ \\
&\quad -\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)}\frac{\rho_{\text{speech}}^{s}}{s!\left(v+1\right)!}\beta_{\text{video}}\left(s,v+1\right)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)}\frac{\rho_{\text{speech}}^{s}}{s!w!}\right) \\
&= \frac{1}{v!w!}\left(\frac{1}{w+1}-\frac{1}{v+1}\right)\times \\
&\quad \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)}\frac{\rho_{\text{speech}}^{s}}{s!}\beta_{\text{video}}\left(s,v+1\right)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)}\frac{\rho_{\text{speech}}^{s}}{s!}\right).
\end{aligned}$$

Note that $\zeta_{v,v}=0$, $v=0,\cdots,v_{\max}-1$, so that the coefficients for ρ^{0} and $\rho^{2(v_{\max}-1)}$ vanish.

Observe that since $\rho_{\text{video}}\geq 0$, a sufficient condition for (5.13) is that all coefficients $\sum_{v+w=k}\zeta_{v,w}\leq 0$, $k=1,\cdots,2v_{\max}-3$. To this end, we will show that $\zeta_{v,w}+\zeta_{w,v}\leq 0$, where we take $v<w$ without loss of generality, i.e.

$$
\begin{aligned}
&\zeta_{v,w} + \zeta_{w,v} \leq 0 \\
&\qquad \Longleftrightarrow \left(\frac{1}{w+1} - \frac{1}{v+1}\right) \times \\
&\qquad\quad \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!} \beta_{\text{video}}\left(s, v+1\right)\right) \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\right) + \\
&\qquad\quad + \left(\frac{1}{v+1} - \frac{1}{w+1}\right) \times \\
&\qquad\quad \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!} \beta_{\text{video}}\left(s, w+1\right)\right) \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\right) \leq 0
\end{aligned}
$$

or, equivalently,

$$
\frac{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!} \beta_{\text{video}}\left(s, v+1\right)}{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}} \geq \frac{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!} \beta_{\text{video}}\left(s, w+1\right)}{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}},
$$

i.e.

$$
\mathbf{E}\left\{\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(v+1)}, v+1\right)\right\} \geq \mathbf{E}\left\{\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(w+1)}, w+1\right)\right\},
$$

where the random variable S_x is distributed as the queue length in a standard Erlang loss model with capacity x and traffic load ρ_{speech}. Observe that effectively we have reduced the inequality $\mathbf{R}_{\text{video}}^{c} \leq \mathbf{R}_{\text{video}}^{t}$ for the SV model to a set of inequalities for a speech-only model, i.e. for the standard Erlang loss model.

To complete the proof, we will show that $\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(v+1)}, v+1\right)$ is almost surely non-increasing in v, for $v = 0, \cdots, v_{\max}-1$. Substituting $y = C - \beta_{\text{video}}^{\min}(v+1)$ we have that

$$
\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(v+1)}, v+1\right) = \beta_{\text{video}}\left(S_y, \frac{C-y}{\beta_{\text{video}}^{\min}}\right),
$$

which we will demonstrate to be almost surely non-decreasing in y, by comparing the above expression for $y, y+\beta_{\text{video}}^{\min} \in \left[0, C-\beta_{\text{video}}^{\min}\right]$, where the lower (upper) bound corresponds with $v = v_{\max} - 1$ ($v = 0$). First observe that the sample paths of the Erlang loss model with capacity y and $y+\beta_{\text{video}}^{\min}$ can readily be compared. Clearly, for an identical input of interarrival times and call lengths it must be that the sample path of the system with capacity $y+\beta_{\text{video}}^{\min}$ is never below that of the system with capacity y. In fact, starting with an empty system, the sample paths coincide until a call is blocked in the system with capacity y. Then, during the period that the system with capacity y is full, it may be that one or more additional calls are admitted to the system with capacity $y+\beta_{\text{video}}^{\min}$. Note that at most $\beta_{\text{video}}^{\min}$ additional calls can be accepted. The sojourn times of the additional calls are independent of the sojourn times of the other calls in the system with capacity $y+\beta_{\text{video}}^{\min}$, which are also present in the system with capacity y. Hence, with probability 1,

$$S_y \leq S_{y+\beta_{\text{video}}^{\min}} \leq S_y + \beta_{\text{video}}^{\min} \text{ and } S_y \leq y.$$

Combining these results with the fact that $y+\beta_{\text{video}}^{\min} \leq C$ and, in general, for $a, b \in \mathbb{R}$ it holds that if $a \geq b > \epsilon$ then $\left(\frac{a-\epsilon}{b-\epsilon}\right) \geq \frac{a}{b}$, implies that

$$\frac{C - S_{y+\beta_{\text{video}}^{\min}}}{C-\left(y+\beta_{\text{video}}^{\min}\right)} \geq \frac{C-\left(S_y+\beta_{\text{video}}^{\min}\right)}{C-\left(y+\beta_{\text{video}}^{\min}\right)} \geq \frac{C-S_y}{C-y},$$

with probability 1. Recall that

$$\beta_{\text{video}}\left(S_y, \frac{C-y}{\beta_{\text{video}}^{\min}}\right) = \min\left\{\beta_{\text{GPRS}}^{\max}, \beta_{\text{video}}^{\min}\frac{C-S_y}{C-y}\right\},$$

so that

$$\beta_{\text{video}}\left(S_{y+\beta_{\text{video}}^{\min}}, \frac{C-\left(y+\beta_{\text{video}}^{\min}\right)}{\beta_{\text{video}}^{\min}}\right) \geq \beta_{\text{video}}\left(S_y, \frac{C-y}{\beta_{\text{video}}^{\min}}\right),$$

with probability 1, which completes the proof. □

Corollary 5.1 *The time-average video throughput is non-increasing in the video traffic load (for $\beta^{\min}_{video} \in \{0, 1, \cdots, C\}$), i.e.*

$$\frac{\partial \mathbf{R}^t_{video}}{\partial \rho_{video}} \leq 0.$$

Proof The proof follows from manipulating the inequality proven in Proposition 5.2, using expressions (5.9) and (5.11), and relating it to the derivative of the time-average video throughput expression (5.11) with respect to ρ_{video}:

$$\begin{aligned}
&\mathbf{R}^c_{\text{video}} \leq \mathbf{R}^t_{\text{video}} \\
&\Longleftrightarrow \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v-1\right) \beta_{\text{video}}\left(s, v\right) \right) \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v\right) \right) + \\
&\qquad - \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v\right) \beta_{\text{video}}\left(s, v\right) \right) \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v-1\right) \right) \leq 0 \\
&\Longleftrightarrow \frac{\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v-1\right) \beta_{\text{video}}\left(s, v\right)}{\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v\right)} + \\
&\qquad - \frac{\left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v\right) \beta_{\text{video}}\left(s, v\right) \right) \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v-1\right) \right)}{\left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \pi\left(s, v\right) \right)^2} \leq 0 \\
&\Longleftrightarrow \frac{\partial \mathbf{R}^t_{\text{video}}}{\partial \rho_{\text{video}}} \leq 0.
\end{aligned}$$

□

5.4.2. V MODEL

All relevant video throughput measures have been explicitly derived for the SV model, including those for the case of unrestricted channel assignments. Therefore, an explicit consideration of the V model would be superfluous, as it is merely a special case of the SV model with $\rho_{\text{speech}} = 0$.

Also the ordering of the different throughput measures is as under the SV model, while we note that the results presented in Proposition 5.2 and Corollary 5.1 are readily proven to hold for arbitrary $\beta_{\text{video}}^{\min}$ in the video-only model, following the same line of proof.

5.4.3. SD MODEL

Consider the SD model with *exponentially* distributed speech call durations and data call sizes. The evolution of the system in the SD model can then be described by an irreducible two-dimensional continuous-time Markov chain $(S(t), D(t))_{t\geq 0}$, with states denoted (s,d). The state space of the Markov chain is given by $\mathbb{S} \equiv \{(s,d) \in \mathbb{N}_0 \times \mathbb{N}_0 : s + d\beta_{\text{data}}^{\min} \leq C$, while its infinitesimal generator $\mathcal{Q}$ is readily specified in terms of the speech and data call arrival and departure rates (analogous to e.g. the SVD model in Section 3.5). The irreducibility of the finite state space Markov chain $(S(t), D(t))_{t\geq 0}$ ensures the existence of a unique probability vector $\boldsymbol{\pi}$ that satisfies the system of global balance equations $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0}$, with $\mathbf{0}$ the vector with all entries zero. The equilibrium distribution is *not* insensitive to the specific form of the speech call duration and data call size distributions. For the Markovian case, the equilibrium distribution can be determined numerically, e.g. by a successive overrelaxation procedure [213].

Using PASTA, the speech and data call blocking probabilities are given by

$$\mathbf{P}_{\text{speech}} = \sum_{d=0}^{d_{\max}(0)} \pi\left(s_{\max}\left(d\right), d\right) \text{ and } \mathbf{P}_{\text{data}} = \sum_{s=0}^{C} \pi\left(s, d_{\max}\left(s\right)\right).$$

In the special case of unrestricted channel assignments to the data service, the speech call blocking probability becomes equal to the Erlang loss probability, as speech traffic does not 'see' data traffic in the absence of data QOS guarantees, while the data call blocking probability becomes zero.

CALL-AVERAGE THROUGHPUT

Compared to other data throughput measures, obtaining explicit expressions for the *call-average* data throughput $\mathbf{R}_{\text{data}}^{c}$ is more involved. We first concentrate on the distribution of the data call sojourn times, conditional on the data call size. For each state $(s,d) \in \mathbb{S}_{\text{data}}^{+} \equiv \{(s,d) \in \mathbb{S} : d > 0\}$ define $\tau_{s,d}(x)$ as the random time it takes to transfer a file of size x, arriving at a given system state (s,d), where d includes the new data call. Define the Laplace-Stieltjes transform of the distribution of $\tau_{s,d}(x)$ by

$$T_{s,d}(\zeta,x) \equiv \mathbf{E}\left\{\exp\left\{-\zeta\tau_{s,d}(x)\right\}\right\}, \ \operatorname{Re}(\zeta) \geq 0, \ (s,d) \in \mathbb{S}^{+}_{\text{data}}$$

and let $\mathbf{T}(\zeta,x) = \left(T_{s,d}(\zeta,x), \ (s,d) \in \mathbb{S}^{+}_{\text{data}}\right)$ be lexicographically ordered in $(s,d) \in \mathbb{S}^{+}_{\text{data}}$.

In an analogous manner as used to determine the conditional expected transfer volumes of video calls in the SV(D) model (see Section 3.6.1), the derivation of an explicit expression for $\mathbf{T}(\zeta,x)$ involves a modified version of the original Markov chain, governed by infinitesimal generator $\mathcal{Q}^{\star}_{\text{data}}$, characterised by the presence of one permanent data call, and with state space $\mathbb{S}^{+}_{\text{data}}$. The data call departure rates in the modified chain reflect the presence of the permanent data call, and are equal to $\mathcal{Q}^{\star}_{\text{data}}\left((s,d)\,;(s,d-1)\right) = \beta_{\text{data}}\left(s,d\right)(d-1)\mu_{\text{data}}$. Denote with $\boldsymbol{\pi}^{\star}_{\text{data}}$ the unique equilibrium distribution of the modified Markov chain and let $\mathcal{B}_{\text{data}} \equiv diag(\beta_{\text{data}}(s,d),$ $(s,d) \in \mathbb{S}^{+}_{\text{data}})$ be the diagonal matrix of data channel assignments, lexicographically ordered in (s,d). Similar to the data transfer time analysis in Section 3.6.1, partition $\mathbb{S}^{+}_{\text{data}}$ into $\mathbb{S}^{+}_{\text{data},0} \equiv \left\{(s,d) \in \mathbb{S}^{+}_{\text{data}} : \beta_{\text{data}}\left(s,d\right) = 0\right\}$ and its complement $\mathbb{S}^{+}_{\text{data},+} \equiv \mathbb{S}^{+}_{\text{data}} \backslash \mathbb{S}^{+}_{\text{data},0}$, and reorder the rows and columns in $\mathcal{Q}^{\star}_{\text{data}}$, $\mathcal{B}_{\text{data}}$, $\boldsymbol{\pi}^{\star}_{\text{data}}$ and $\mathbf{T}(\zeta,x)$ in accordance with the introduced state space partitioning, in order to allow the partitioning

$$\mathcal{Q}^{\star}_{\text{data}} = \begin{bmatrix} \mathcal{Q}^{\star}_{++} & \mathcal{Q}^{\star}_{+0} \\ \mathcal{Q}^{\star}_{0+} & \mathcal{Q}^{\star}_{00} \end{bmatrix}, \ \mathcal{B}_{\text{data}} = \begin{bmatrix} \mathcal{B}_{+} & \mathcal{O} \\ \mathcal{O} & \mathcal{O} \end{bmatrix},$$

and

$$\boldsymbol{\pi}^{\star}_{\text{data}} = \left(\boldsymbol{\pi}^{\star}_{\text{data},0}, \boldsymbol{\pi}^{\star}_{\text{data},+}\right), \ \mathbf{T}(\zeta,x) = \left(\mathbf{T}_{0}(\zeta,x), \mathbf{T}_{+}(\zeta,x)\right),$$

where we omit the ‘data’ subscript in the submatrices of $\mathcal{Q}^{\star}_{\text{data}}$ and $\mathcal{B}_{\text{data}}$ for enhanced readability. We note that in case $\beta^{\min}_{\text{data}} > 0$, this implies that $\mathbb{S}^{+}_{\text{data},0} = \emptyset$, leading to a slightly simplified analysis (see [173, Section 4.2]).

As shown in [173, Section 4.4], for $x \geq 0$ and $\mathrm{Re}(\zeta) \geq 0$, a closed-form expression for $\mathbf{T}(\zeta, x)$ is given by

$$\mathbf{T}_0(\zeta, x) = -\left(\mathcal{Q}^\star_{00} - \zeta\mathcal{I}\right)^{-1} \mathcal{Q}^\star_{0+}\mathbf{T}_+(\zeta, x),$$

and

$$\mathbf{T}_+(\zeta, x) = \exp\left\{x\mathcal{B}_+^{-1}\left(\mathcal{Q}^\star_{++} - \mathcal{Q}^\star_{+0}\left(\mathcal{Q}^\star_{00} - \zeta\mathcal{I}\right)^{-1}\mathcal{Q}^\star_{0+} - \zeta\mathcal{I}\right)\right\}\mathbf{1}.$$

The conditional expected throughput $\mathbf{R}^c_{\mathrm{data}}(s, d, x)$ of a data call admitted to the system in state (s, d) and with a given size x is given by

$$\begin{aligned}
\mathbf{R}^c_{\mathrm{data}}(s, d, x) &= r_{\mathrm{data}}\mathbf{E}\left\{\frac{x}{\tau_{s,d}\left(x\right)}\right\} \\
&= r_{\mathrm{data}}\int\limits_{\tau=0}^{\infty}\frac{x}{\tau}d\Phi_{s,d,x}\left(\tau\right) \\
&= r_{\mathrm{data}}x\int\limits_{\tau=0}^{\infty}\left(\int\limits_{\zeta=0}^{\infty}\exp\left\{-\zeta\tau\right\}d\zeta\right)d\Phi_{s,d,x}\left(\tau\right) \\
&= r_{\mathrm{data}}x\int\limits_{\zeta=0}^{\infty}\left(\int\limits_{\tau=0}^{\infty}\exp\left\{-\zeta\tau\right\}d\Phi_{s,d,x}\left(\tau\right)\right)d\zeta \\
&= r_{\mathrm{data}}x\int\limits_{\zeta=0}^{\infty}T_{s,d}(\zeta, x)d\zeta,
\end{aligned}$$

where $\Phi_{s,d,x}\left(\tau\right)$ denotes the cumulative distribution function of $\tau_{s,d}\left(x\right)$ given data call size x and system state (s, d) upon the considered data call's admission. Deconditioning on the system state (s, d) upon admission yields

$$\mathbf{R}^c_{\mathrm{data}}(x) = \sum_{(s,d)\in\mathbb{S}^+_{\mathrm{data}}}\left(\frac{\pi(s, d-1)}{\sum\limits_{(s',d')\in\mathbb{S}^+_{\mathrm{data}}}\pi(s', d'-1)}\right)\mathbf{R}^c_{\mathrm{data}}(s, d, x),$$

while subsequently deconditioning on the exponentially distributed data call size x gives the call-average data throughput:

$$\mathbf{R}^c_{\text{data}} = \mu_{\text{data}} \sum_{(s,d)\in\mathbb{S}^+_{\text{data}}} \left(\frac{\pi(s,d-1)}{\sum\limits_{(s',d')\in\mathbb{S}^+_{\text{data}}} \pi(s',d'-1)} \right) \int\limits_{x=0}^{\infty} \exp\left(-\mu_{\text{data}} x\right) \mathbf{R}^c_{\text{data}}(s,d,x)dx.$$

TIME-AVERAGE THROUGHPUT

By analogy with the derivation of the time-average video throughput in (5.11), the *time-average* data throughput is obtained as

$$\mathbf{R}^t_{\text{data}} = r_{\text{data}} \sum_{(s,d)\in\mathbb{S}^+_{\text{data}}} \left(\frac{\pi(s,d)}{\sum\limits_{(s',d')\in\mathbb{S}^+_{\text{data}}} \pi(s',d')} \right) \beta_{\text{data}}\left(s,d\right).$$

Since the equilibrium distribution can only be numerically obtained, the above expression does not simplify for the special case of unrestricted channel assignments.

EXPECTED INSTANTANEOUS THROUGHPUT

Similar to the derivation of the corresponding measure (5.12) for the SV model, the expected *instantaneous* data throughput is given by

$$\mathbf{R}^i_{\text{data}} = r_{\text{data}} \sum_{(s,d)\in\mathbb{S}^+_{\text{data}}} \left(\frac{\pi(s,d-1)}{\sum\limits_{(s',d')\in\mathbb{S}^+_{\text{data}}} \pi(s',d'-1)} \right) \beta_{\text{data}}\left(s,d\right).$$

RATIO THROUGHPUT MEASURE

The *ratio* of the expected data call size and the expected data call sojourn time is equal to

$$\mathbf{R}^r_{\text{data}} = \left(\frac{r_{\text{data}}}{\mu_{\text{data}}}\right) \Bigg/ \left(\frac{\sum\limits_{(s,d)\in\mathbb{S}} d\pi(s,d)}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})}\right) = r_{\text{data}} \frac{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}{\sum\limits_{(s,d)\in\mathbb{S}} d\pi(s,d)},$$

where Little's formula (see e.g. [213]) is applied to derive the expected data call sojourn time.

CALL-AVERAGE STRETCH

Using

$$\mathbf{E}\left\{\frac{\tau_{s,d}(x)}{x}\right\} = -\frac{1}{x}\frac{\partial}{\partial\zeta}T_{s,d}(\zeta,x)\bigg|_{\zeta=0},$$

with $T_{s,d}(\zeta,x)$ as defined above, the expected (call-average) data stretch is given by

$$\begin{aligned}\mathbf{S}_{\text{data}} &= C\,\mathbf{E}\left\{\frac{\tau_{s,d}(x)}{x}\right\} = \\ &= -C\mu_{\text{data}} \sum_{(s,d)\in\mathbb{S}^+_{\text{data}}} \left(\frac{\pi(s,d-1)}{\sum\limits_{(s',d')\in\mathbb{S}^+_{\text{data}}} \pi(s,d-1)}\right) \times \\ &\quad \times \left\{\int\limits_{x=0}^{\infty} \frac{1}{x}\exp(-\mu_{\text{data}}x)\left(\frac{\partial}{\partial\zeta}T_{s,d}(\zeta,x)\bigg|_{\zeta=0}\right)dx\right\},\end{aligned}$$

conform the definition given by (5.5), noting that in the above analysis the data call size x is expressed in units of r_{data} kbits (see also Section 5.3.1).

COMPARISON OF MEASURES

The throughput measures derived for the SD model do not allow an analytical comparison, as we have been unable to derive explicit expressions. A numerical comparison is presented in Section 5.5.

5.4.4. D MODEL

The D model is a special case of the SD model with $\rho_{\text{speech}} = 0$. Moreover, the D model is equivalent to the $M/G/1/d_{\max}/GPS$ queueing model with state-dependent aggregate service rates given by $dr_{\text{data}}\beta_{\text{data}}(d) = dr_{\text{data}} \min\{C/d, \beta_{\text{GPRS}}^{\max}\}$, see [54]. For this model, the equilibrium distribution is known to be *insensitive* to the specific form of the data call size distribution, and is given by

$$\pi(d) = \frac{\left(\rho_{\text{data}}^{\star}\right)^{d} \phi(d)}{\sum\limits_{d'=0}^{d_{\max}} \left(\rho_{\text{data}}^{\star}\right)^{d'} \phi(d')} \text{ with } \phi(d) \equiv \left(\prod_{d'=1}^{d} \frac{d'\beta_{\text{data}}(d')}{C}\right)^{-1},$$

$d = 0, \cdots, d_{\max}$, where $\rho_{\text{data}}^{\star} \equiv \rho_{\text{data}}/C$ denotes the normalised data traffic load and $\phi(0) \equiv 1$ by convention. For the special case of unrestricted channel assignments, the D model reduces to the standard $M/G/1/PS$ queueing model, which has a geometric equilibrium distribution:

$$\widetilde{\pi}(d) = \left(1 - \rho_{\text{data}}^{\star}\right)\left(\rho_{\text{data}}^{\star}\right)^{d},\ d \geq 0,$$

requiring $\rho_{\text{data}}^{\star} < 1$ for stability.

Using PASTA, the data call blocking probability is equal to

$$\mathbf{P}_{\text{data}} = \pi(d_{\max}),$$

while it is equal to zero in the case of unrestricted channel assignments.

CALL-AVERAGE THROUGHPUT

The *call-average* data throughput is not insensitive, as will be demonstrated in Section 5.5, and we therefore concentrate on the case of exponentially distributed data call sizes. We first derive a closed-form expression for $\mathbf{T}(\zeta, x) \equiv (T_d(\zeta, x),\ d = 1, \cdots, d_{\max})$ with $T_d(\zeta, x)$ the Laplace-Stieltjes transform of the distribution of $\tau_d(x)$, i.e. the random sojourn time of a data call of size x admitted to the system in the presence of $d-1$ other data calls. Recall that x is expressed in the nominal sojourn time (in seconds). By analogy with the similar analysis presented for the SD model, $\mathcal{B}_{\text{data}}$ is the diagonal matrix of channel assignments and $\mathcal{Q}^{\star}_{\text{data}}$ is the infinitesimal generator corresponding the D model's modified Markov chain with one permanent data call. In this data-only model, $\beta_{\text{data}}(d) > 0$ for all $d \geq 1$, so that no partitioning of $\mathbf{T}(\zeta, x)$ is required. As a specific instance of the result given in Section 5.4.3, for $x \geq 0$ and $\text{Re}(\zeta) \geq 0$, $\mathbf{T}(\zeta, x)$ is given by the closed-form expression

$$\mathbf{T}(\zeta, x) = \exp\left\{x\mathcal{B}^{-1}_{\text{data}}\left(\mathcal{Q}^{\star}_{\text{data}} - \zeta\mathcal{I}\right)\right\}\mathbf{1}.$$

By analogy with the analysis for the SD model, expressions for the conditional expected throughput measures $\mathbf{R}^{c}_{\text{data}}(d, x)$ and $\mathbf{R}^{c}_{\text{data}}(x)$ are readily derived. We limit ourselves here to stating the (unconditional) call-average data throughput:

$$\mathbf{R}^{c}_{\text{data}} = \mu_{\text{data}} \sum_{d=1}^{d_{\max}} \left(\frac{\pi(d-1)}{\sum\limits_{d'=1}^{d_{\max}} \pi(d'-1)} \right) \int\limits_{x=0}^{\infty} \exp\left(-\mu_{\text{data}} x\right) \left(r_{\text{data}} x \int\limits_{\zeta=0}^{\infty} T_d(\zeta, x) d\zeta \right) dx.$$

For the case of unrestricted channel assignments, $\widetilde{\mathbf{R}}^{c}_{\text{data}}(x)$ can be obtained using the following closed-form expression for the deconditioned Laplace-Stieltjes transform $\widetilde{T}(\zeta, x)$ as derived in [52]:

$$\begin{aligned}
\widetilde{T}(\zeta, x) \quad &\equiv \quad \mathbf{E}\left\{\exp\left\{-\zeta\tau(x)\right\}\right\} = \sum_{d=1}^{\infty} \left(\frac{\pi(d-1)}{\sum\limits_{d'=1}^{\infty} \pi(d'-1)} \right) \widetilde{T}_d(\zeta, x) \\
&= \quad \frac{\left(1 - \rho^{\star}_{\text{data}}\right)\left(1 - \rho^{\star}_{\text{data}} r^2\right) \exp\left\{-\left(\lambda_{\text{data}}(1-r) + \zeta\right) x\right\}}{\left(1 - \rho^{\star}_{\text{data}} r\right)^2 - \rho^{\star}_{\text{data}} (1-r)^2 \exp\left\{-\mu x \left(1 - \rho^{\star}_{\text{data}} r^2\right)/r\right\}},
\end{aligned}$$

with $\text{Re}(\zeta) \geq 0$ and r given by

$$r = \frac{\left(\lambda_{\text{data}} + \mu_{\text{data}} + \zeta\right) - \sqrt{\left(\lambda_{\text{data}} + \mu_{\text{data}} + \zeta\right)^2 - 4\lambda_{\text{data}}\mu_{\text{data}}}}{2\lambda_{\text{data}}},$$

so that the conditional expected (call-average) data throughput is given by

$$\begin{aligned}
\widetilde{\mathbf{R}}^{c}_{\text{data}}(x) &= \sum_{d=1}^{\infty} \pi(d-1)\left(r_{\text{data}}x \int\limits_{\zeta=0}^{\infty} \widetilde{T}_{d}(\zeta,x)d\zeta \right) \\
&= r_{\text{data}}x \int\limits_{\zeta=0}^{\infty} \left(\sum_{d=1}^{\infty} \pi(d-1)\widetilde{T}_{d}(\zeta,x) \right) d\zeta = r_{\text{data}}x \int\limits_{\zeta=0}^{\infty} \widetilde{T}(\zeta,x)d\zeta \\
&= r_{\text{data}}x \int\limits_{\zeta=0}^{\infty} \frac{(1-\rho)\left(1-\rho r^{2}\right)\exp\left\{-\left(\lambda\left(1-r\right)+\zeta\right)x\right\}}{\left(1-\rho r\right)^{2}-\rho\left(1-r\right)^{2}\exp\left\{-\mu x\left(1-\rho r^{2}\right)/r\right\}} d\zeta.
\end{aligned}$$

TIME-AVERAGE THROUGHPUT

The *time-average* data throughput is given by

$$\mathbf{R}^{t}_{\text{data}} = r_{\text{data}} \sum_{d=1}^{d_{\max}} \left(\frac{\pi(d)}{\sum\limits_{d'=1}^{d_{\max}} \pi(d')} \right) \beta_{\text{data}}(d),$$

while in the case of unrestricted channel assignments, the time-average data throughput is given by

$$\begin{aligned}
\widetilde{\mathbf{R}}^{t}_{\text{data}} &= r_{\text{data}} \sum_{d=1}^{\infty} \left(\frac{\left(1-\rho^{\star}_{\text{data}}\right)\left(\rho^{\star}_{\text{data}}\right)^{d}}{\sum\limits_{d'=1}^{\infty}\left(1-\rho^{\star}_{\text{data}}\right)\left(\rho^{\star}_{\text{data}}\right)^{d'}} \right) \frac{C}{d} \\
&= r_{\text{data}} C \left(\frac{1-\rho^{\star}_{\text{data}}}{\rho^{\star}_{\text{data}}} \right) \sum_{d=1}^{\infty} \left(\frac{\left(\rho^{\star}_{\text{data}}\right)^{d}}{d} \right) \\
&= r_{\text{data}} C \left(\frac{1-\rho^{\star}_{\text{data}}}{\rho^{\star}_{\text{data}}} \right) \ln\left(\frac{1}{1-\rho^{\star}_{\text{data}}} \right),
\end{aligned}$$

requiring $\rho^{\star}_{\text{data}} < 1$ for stability. Note that due to the insensitivity of the equilibrium distribution, these expressions for the time-average throughput are also insensitive to the specific form of the data call size distribution

EXPECTED INSTANTANEOUS THROUGHPUT

The expected *instantaneous* data throughput is given by

$$\mathbf{R}^{i}_{\text{data}} = r_{\text{data}} \sum_{d=1}^{d_{\max}} \left(\frac{\pi(d-1)}{\sum\limits_{d'=1}^{d_{\max}} \pi(d'-1)} \right) \beta_{\text{data}}\left(d\right). \tag{5.14}$$

In the special case of unrestricted channel assignments, the expected instantaneous data throughput is equal to the time-average data throughput:

$$\widetilde{\mathbf{R}}^{i}_{\text{data}} = r_{\text{data}} C \left(\frac{1-\rho^{\star}_{\text{data}}}{\rho^{\star}_{\text{data}}} \right) \ln \left(\frac{1}{1-\rho^{\star}_{\text{data}}} \right),$$

requiring $\rho^{\star}_{\text{data}} < 1$ for stability. Once again, the above expressions for the expected instantaneous throughputs inherit the insensitivity property of the equilibrium distribution.

RATIO THROUGHPUT MEASURE

The *ratio* of the expected data call size and the expected data call sojourn time is equal to

$$\mathbf{R}^{r}_{\text{data}} = r_{\text{data}} \frac{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}{\sum\limits_{d=0}^{d_{\max}} d\pi(d)},$$

while in the case of unrestricted channel assignments we have

$$\widetilde{\mathbf{R}}^{r}_{\text{data}} = r_{\text{data}} C \left(1-\rho^{\star}_{\text{data}}\right),$$

requiring $\rho^{\star}_{\text{data}} \leq 1$. Both expressions are insensitive to the data call size distribution aside from its mean.

CALL-AVERAGE STRETCH

The call-average *stretch* is given by

$$\begin{aligned}\mathbf{S}_{\text{data}} &= \mathbf{E}\left\{\mathbf{S}_{\text{data}}(x)\right\} = C\,\mathbf{E}\left\{\frac{\mathbf{T}_{\text{data}}(x)}{x}\right\} \\ &= C\,\mathbf{E}\left\{\frac{1}{x}\left(x\frac{\sum\limits_{d=0}^{d_{\max}} d\pi(d)}{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}\right)\right\} = \frac{\sum\limits_{d=0}^{d_{\max}} d\pi(d)}{\rho^{\star}_{\text{data}}(1-\mathbf{P}_{\text{data}})},\end{aligned}$$

using the known linearity in x of the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call of size x [54, 213]. The call-average stretch for the case of unrestricted channel assignments is readily derived to be equal to

$$\widetilde{\mathbf{S}}_{\text{data}} = \frac{1}{1-\rho^{\star}_{\text{data}}},$$

requiring $\rho^{\star}_{\text{data}} < 1$ for stability. Note that the effect of the channel rate r_{data} is captured only in the definition of the data traffic load $\rho^{\star}_{\text{data}}$.

COMPARISON OF MEASURES

A number of relations between the different throughput measures derived above can be summarised.

Proposition 5.3

$$\mathbf{R}^{c}_{data} \geq \mathbf{R}^{r}_{data}. \tag{5.15}$$

Proof The result is a straightforward extension of the equivalent result given in [127] for the case of unrestricted channel assignments. Applying Jensen's inequality (see e.g. [200]) with convex mapping $\psi(x) \equiv 1/x$:

$$\mathbf{R}^{c}_{\text{data}} = r_{\text{data}}\mathbf{E}\left\{\psi\left(\frac{\mathbf{T}_{\text{data}}(x)}{x}\right)\right\}$$

$$
\begin{aligned}
&\geq \quad r_{\text{data}}\psi\left(\mathbf{E}\left\{\frac{\mathbf{T}_{\text{data}}\,(x)}{x}\right\}\right) = r_{\text{data}}\left(\mathbf{E}\left\{\frac{1}{x}\left(x\frac{\sum\limits_{d=0}^{d_{\max}} d\pi(d)}{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}\right)\right\}\right)^{-1} \\
&= \quad r_{\text{data}}\frac{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}{\sum\limits_{d=0}^{d_{\max}} d\pi(d)} = \mathbf{R}^{r}_{\text{data}}.
\end{aligned}
$$

□

We further adopt the following result for the case of unrestricted channel assignments and deterministic data call sizes.

Proposition 5.4 (Kherani and Kumar [127]**)** *In case of deterministic data call sizes, the following inequality holds:*

$$\widetilde{\mathbf{R}}^{t}_{data} > \widetilde{\mathbf{R}}^{c}_{data}. \tag{5.16}$$

Lastly, the explicitly derived expressions above reveal that, *only* for the case of unrestricted channel assignments, the time-average throughput is equal to the expected instantaneous throughput:

$$\widetilde{\mathbf{R}}^{t}_{\text{data}} = \widetilde{\mathbf{R}}^{i}_{\text{data}},$$

while in general it holds that

$$\mathbf{R}^{r}_{\text{data}}\mathbf{S}_{\text{data}} = \widetilde{\mathbf{R}}^{r}_{\text{data}}\widetilde{\mathbf{S}}_{\text{data}} = r_{\text{data}}C.$$

5.5. NUMERICAL RESULTS

Whereas the previous section contained an analytical evaluation and comparison, we now present the results from a set of numerical experiments, carried out in order to provide further insight in the throughput performance of elastic (video or data) calls in a system with a fixed or varying service capacity. The applied system and traffic parameter settings are summarised in Table 5.1. We argue however that the revealed qualitative trends are unaffected by the actual parameter settings and thus apply to contexts other than that of a GSM/GPRS cell equally well. The number of traffic channels C in the integrated services SV/SD models is based on a cell with 22 traffic channels (corresponding to 3 GSM frequencies *minus* 2 control channels). The capacity selected for the single service V/D models is equal to the average number of idle traffic channels in the SV/SD models, i.e. $22 - \rho_{\text{speech}}\,(1 - \mathbf{P}_{\text{speech}})$, where ρ_{speech} is chosen such the corresponding speech call blocking probability is 1%. The speech call durations are exponentially distributed, in correspondence with the numerical experiments of Chapter 4 in order to allow a comparison of the obtained insights. An average call duration of 50 seconds is assumed for both the speech and video service. The average data file transfer is set at 320 kbits, which normalises to the given expected duration of μ_{data}^{-1} seconds. The video (data) bit rate per traffic channel is set to 13.4 (9.05) kbits/s, based on an assumed GPRS coding scheme CS-2 (CS-1). The video and data traffic loads are varied between 0 and the applicable value of C. Potential practical upper bounds on the channel assignment are disregarded. In the conditional throughput analyses for the V/D models, the minimum QOS requirements are varied within the range $[0, C]$, so that corresponding CAC thresholds between 1 and ∞ are considered, while no such restrictions are imposed for the unconditional throughput analyses.

Table 5.1 Summary of the parameter settings assumed for the numerical experiments, based on the chosen context of a single cell in a GSM/GPRS network.

	SV model	V model	SD model	D model
C	22	8.486	22	8.486
μ_{speech}^{-1}	50 seconds	-	50 seconds	-
ρ_{speech}	13.651 Erlang	-	13.651 Erlang	-
μ_{video}^{-1}	50 seconds	50 seconds	-	-
ρ_{video}	$\in (0, C)$	$\in (0, C)$	-	-
r_{video}	13.4 kbits/s	13.4 kbits/s	-	-
$\beta_{\text{video}}^{\min}$	0 channels	$\in [0, C]$ channels	-	-
μ_{data}^{-1}	-	-	35.359 seconds	35.359 seconds
ρ_{data}	-	-	$\in (0, C)$	$\in (0, C)$
r_{data}	-	-	9.05 kbits/s	9.05 kbits/s
$\beta_{\text{data}}^{\min}$	-	-	0 channels	$\in [0, C]$ channels
$\beta_{\text{GPRS}}^{\max}$	C	C	C	C

The next subsection contains a numerical evaluation of the conditional call-average throughput in the V and D models as a function of the (exponentially distributed) elastic call size, the number of competing elastic calls found upon admission and the CAC threshold. Subsequently, an extensive numerical investigation is presented of the various (unconditional) throughput measures as a function of the elastic traffic load, considering different elastic call size distributions where relevant. As the results will demonstrate, the expected instantaneous throughput is the only throughput measure that closely approximates the call-average throughput for all considered scenarios.

5.5.1. CONDITIONAL THROUGHPUT RESULTS

We now present the results of the numerical conditional throughput analyses that have been carried out for the single service V and D models, respectively.

V MODEL

Figure 5.1 shows the conditional call-average video throughputs (in kbits/s) for the case of exponentially distributed video call durations and $\rho_{\mathrm{video}} = \frac{1}{2}C = 11$. A logarithmic scale is used for the video call duration τ (expressed in seconds). The results in the left chart assume a CAC threshold of $v_{\max} = 10$, which is achieved by setting $\beta_{\mathrm{video}}^{\min} \in (0.7715, 0.8486]$, and leads to a video call blocking probability of $\mathbf{P}_{\mathrm{video}} = 0.0075$. The depicted curve for $\mathbf{R}_{\mathrm{video}}^{c}(v, \tau)$ is obtained using a special case of the result presented in (5.8), i.e. without speech traffic. As $\tau \downarrow 0$, the call-average throughputs conditional on the system state v upon admission approach $r_{\mathrm{video}}\beta_{\mathrm{video}}(v) = 113.7023/v$. As τ increases the impact of the system state upon admission vanishes and for each v the call-average throughput converges towards the time-average video throughput in a system with one permanent video call, which was seen to be equal to $\mathbf{R}_{\mathrm{video}}^{c}$, i.e. the call-average video throughput in the original model without a permanent video call, h.l. equal to 26.6132. Observe that for low (high) v, convergence is from above (below), in accordance with intuition.

The right chart shows $\mathbf{R}_{\mathrm{video}}^{c}(\tau)$ for $\beta_{\mathrm{video}}^{\min} \in [0, C]$ and hence $v_{\max} \in \{1, 2 \cdots, \infty\}$. The corresponding video call blocking probabilities are as follows:

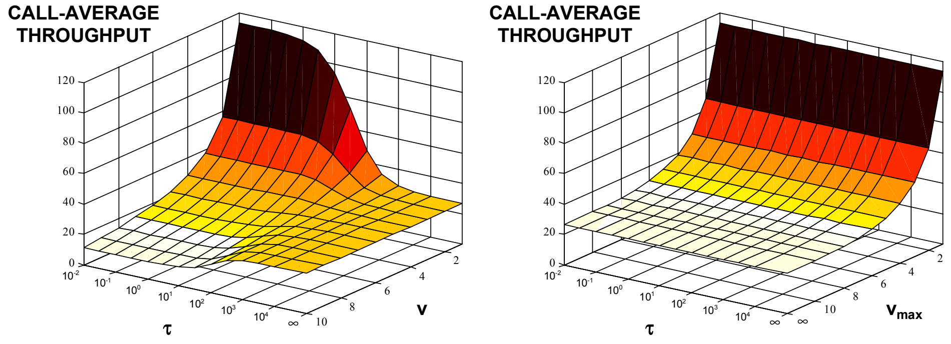


Figure 5.1 Conditional expected throughput performance in the V model. The left (right) chart shows the call-average throughput of a tagged video call as a function of its duration τ and the number of video calls v found upon admission (the CAC threshold $v_{\max}$).

$v_{\max}$	1	2	3	4	5	10	∞
$\mathbf{P}_{\text{video}}$	0.8093	0.6319	0.4719	0.3336	0.2206	0.0075	0.0000

Observe that indeed, as proven above, $\mathbf{R}^{c}_{\text{video}}(\tau)$ is independent of the video call duration τ, which reflects the equivalence of the expected instantaneous throughput and call-average throughput measures. For $v_{\max} = 1$, the call-average video throughput is trivially equal to the aggregate service rate $r_{\text{video}}C = 113.7023$, while for $v_{\max} \to \infty$ the conditional video throughput decays exponentially to

$$r_{\text{video}}C\left(\frac{1-\exp\left(-\rho_{\text{video}}\right)}{\rho_{\text{video}}}\right) = 26.4149.$$

Note that the case for $v_{\max} = 10$ is identical to the converged values in the left chart (for $\tau \to \infty$).

D MODEL

Figure 5.2 shows the conditional call-average data throughputs in the D model, for the case of exponentially distributed data call sizes and $\rho_{\text{data}} = \frac{1}{2}C = 11$ ($\rho^{\star}_{\text{data}} = 0.5$). Equivalent to the above experiment for the V model, the results for $\mathbf{R}^{c}_{\text{data}}(d,x)$

(with x expressed in nominal transfer seconds, as explained in Section 5.3) in the left chart assume a CAC threshold of $d_{\max} = 10$, which is achieved by setting $\beta_{\text{data}}^{\min} \in (0.7715, 0.8486]$. At the considered data traffic load, the selected CAC threshold causes virtually no data call blocking. The profile of the left chart is very similar to that of the left chart in Figure 5.1: $\lim_{x\downarrow 0} \mathbf{R}_{\text{data}}^{c}(d,x)$ is given by the instantaneous throughput $r_{\text{data}}\beta_{\text{data}}(d) = 76.7915/d$, while $\lim_{x\to\infty} \mathbf{R}_{\text{data}}^{c}(d,x)$ is independent of d and given by the time-average data throughput in a data-only system with one permanent call, readily derived to be

$$r_{\text{data}}C\frac{\left(1-\rho_{\text{data}}^{\star}\right)\left(1-\left(\rho_{\text{data}}^{\star}\right)^{d_{\max}}\right)}{\left(1-\left(\rho_{\text{data}}^{\star}\right)^{d_{\max}+1}\right)-\left(d_{\max}+1\right)\left(\rho_{\text{data}}^{\star}\right)^{d_{\max}}\left(1-\rho_{\text{data}}^{\star}\right)} = 38.5843. \qquad (5.17)$$

In contrast with the V model, in the D model the time-average throughput in the modified Markov chain with one permanent data call is *not* equal to the call-average throughput in the original Markov chain.

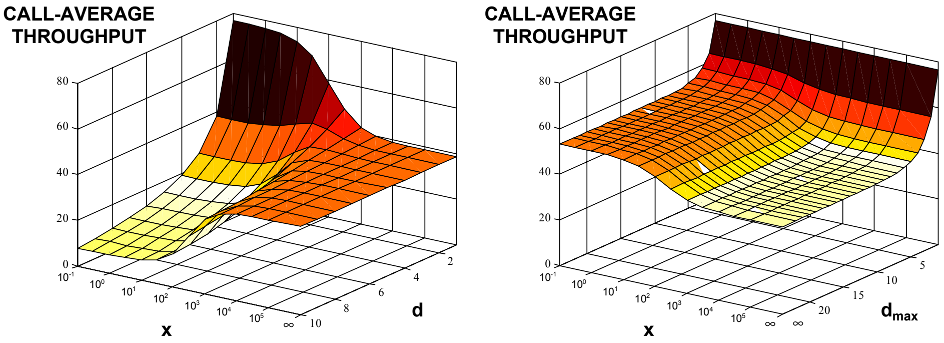


Figure 5.2 Conditional expected throughput performance in the D model. The left (right) chart shows the call-average throughput of a tagged data call as a function of its size x and the number of data calls d found upon admission (the CAC threshold $d_{\max}$).

The right chart shows $\mathbf{R}_{\text{data}}^{c}(x)$ for various CAC thresholds $d_{\max} \in \{1, 2\cdots, \infty\}$, with the corresponding data call blocking probabilities given by

$d_{\max}$	1	2	3	4	5	10	∞
$\mathbf{P}_{\text{data}}$	0.3333	0.1429	0.0667	0.0323	0.0159	0.0005	0.0000

In the trivial case of $d_{\max} = 1$, the call-average data throughput is equal to the aggregate service rate $r_{\text{data}}C = 76.7915$, independent of the data call size x. As $d_{\max}$ increases, not only does $\mathbf{R}^c_{\text{data}}(x)$ decrease due to an increased carried data traffic load and hence a greater competition for resources, it is also no longer independent of x. For a given CAC threshold of $d_{\max}$, $\mathbf{R}^c_{\text{data}}(x)$ decreases from the corresponding expected instantaneous data throughput $\mathbf{R}^i_{\text{data}}$ (cf. expression (5.14)) to the expected time-average data throughput in the associated modified Markov chain with one permanent data call (cf. expression (5.17)). Observe that the expected instantaneous throughput is an upper bound for the call-average throughput. Unlike in the V model, in the D model small calls experience a higher throughput than large calls. It is stressed, however, that the expected sojourn time is proportional to the data call size, so that the expected stretch is insensitive to the data call size. The potential confusion is due to the fact that the reciprocal of the expectation of a random variable is generally unequal to the expectation of the reciprocal of that random variable.

5.5.2. UNCONDITIONAL THROUGHPUT RESULTS

The remainder of this numerical section concentrates on the unconditional throughput as a function of the elastic traffic load, with a principal focus on the proximity of the various throughput measures in the different PS models.

(S)V MODEL

Figure 5.3 depicts the various (unconditional) throughput performance measures as a function of the normalised elastic traffic load. In all considered cases channel assignment restrictions have been imposed on the elastic services. The left chart covers both the SV and the V models, for which all throughput measures are identical for any given normalised video traffic load $\rho^{\star}_{\text{video}} \equiv \rho_{\text{video}}/C$, with C appropriately chosen in each model (see Table 5.1). The chart reveals both the demonstrated equality of $\mathbf{R}^c_{\text{video}}$, $\mathbf{R}^i_{\text{video}}$ and $\mathbf{R}^r_{\text{video}}$, and the proven ordering of $\mathbf{R}^t_{\text{video}} \geq \mathbf{R}^c_{\text{video}}$. It can be observed from the numerical results that $\mathbf{R}^t_{\text{video}}$ may exceed $\mathbf{R}^c_{\text{video}}$ by more than 36%.

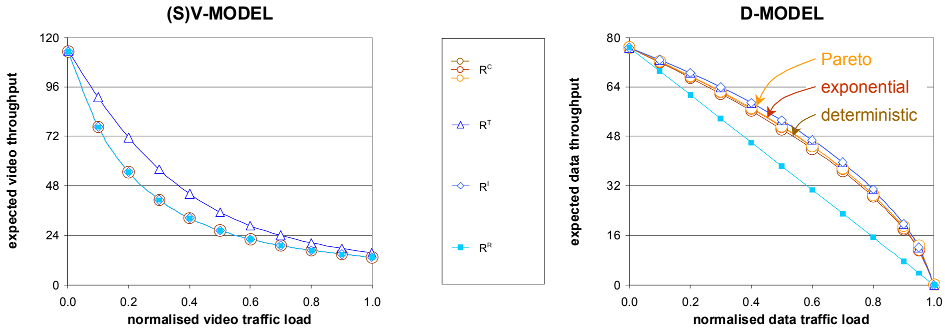


Figure 5.3 Comparison of different throughput measures in the SV, V and D models. The (insensitive) throughput measures in the left chart are identical for the SV and V models, given an appropriately normalised video traffic load. The right chart depicts for the D model the insensitive $\mathbf{R}^t_{\text{data}}$, $\mathbf{R}^i_{\text{data}}$ and $\mathbf{R}^r_{\text{data}}$ measures, along with the sensitive $\mathbf{R}^c_{\text{data}}$ measure for three distinct data call size distributions.

D MODEL

The right chart concentrates on the D model. Since (only) the call-average throughput measure $\mathbf{R}^c_{\text{data}}$ is sensitive to the data call size distribution and no explicit expression could be derived, three distinct curves have been obtained via dynamic simulations for deterministic (zero variance), exponential and Pareto (with shape parameter $\alpha = 1.35$: infinite variance) data call size distributions. Sufficient numerical accuracy is ensured in the simulation experiment, indicated by a relative precision of the constructed 95% confidence intervals that is no worse than 5%. Observe that the call-average throughput is higher for more variable data call sizes, as also observed in [127], although the discrepancies are extremely small. This is probably due to the fact that a more variable data call size distribution features a relatively large number of small data calls, which appear to experience higher throughputs than large data calls (cf. the right chart of Figure 5.2). Recall that in Chapter 4 of this monograph we presented an extensive sensitivity analysis of the data QOS with respect to the data call size variability in the SD model, concentrating on expected sojourn times rather than throughput performance, where an analogous trend was observed and analytically supported.

As shown in Section 5.4, the insensitive time-average and expected instantaneous throughput measures are identical, and appear to offer a very good, only slightly overestimating (cf. (5.16)), approximation for the call-average throughput. Finally, $\mathbf{R}^r_{\text{data}}$ significantly underestimates the call-average throughput (cf. (5.15)), for high data traffic loads even by a factor exceeding 2.

SD MODEL

For the SD model all the throughput measures are more or less sensitive to the data call size distribution, so that for reasons of clarity the numerical results are presented in the two separate charts of Figure 5.4 (for each marker in the legend, the left (right) throughput measure listed is associated with the left (right) chart). In all cases, observe again that a more variable data call size distribution appears to lead to higher expected throughputs, which is in agreement with the sojourn time results of Chapter 4. In this data model with varying service capacity, both the time-average throughput ($\mathbf{R}^t_{\text{data}}$) and the ratio of the expected data call size and the expected sojourn time ($\mathbf{R}^r_{\text{data}}$) are significantly lower than the call-average throughput ($\mathbf{R}^c_{\text{data}}$), in particular for lower data traffic loads. In contrast, the expected instantaneous throughput ($\mathbf{R}^i_{\text{data}}$) remains to be a very good and fairly insensitive approximation for $\mathbf{R}^c_{\text{data}}$, across the entire range of data traffic loads. The *slight over*estimation of the call-average throughput seems to be not significant enough to lead to perilously loose Call Admission Control schemes or planning guidelines.

Comparing the throughput results for the D and SD models, observe that the call-average data throughput appears to be fairly insensitive to the variability of the available capacity, as also observed in [61] (recall that for the SV and V models, the call-average video throughputs were identical). Only for heavy data traffic loads, the call-average data throughput is non-negligibly higher for the fixed capacity D model.

In order to get a better grasp on the large discrepancy between e.g. the time- and call-average data throughputs in the SD model, the left chart of Figure 5.5 shows the time-average data throughput versus the normalised data traffic load for various degrees of acceleration of the speech call arrival and departure process. Keeping ρ_{speech} fixed at 13.651 Erlang, we multiply both λ_{speech} and μ_{speech} by the acceleration factor $\vartheta \in \{1, 10, 100, \infty\}$. The case of $\vartheta = 1$ refers to the original model and the associated curve is identical to the one for $\mathbf{R}^t_{\text{data}}$ in Figure 5.4 (left chart). At the other extreme, in the case of $\vartheta \to \infty$ the speech calls arrive and depart so quickly, that

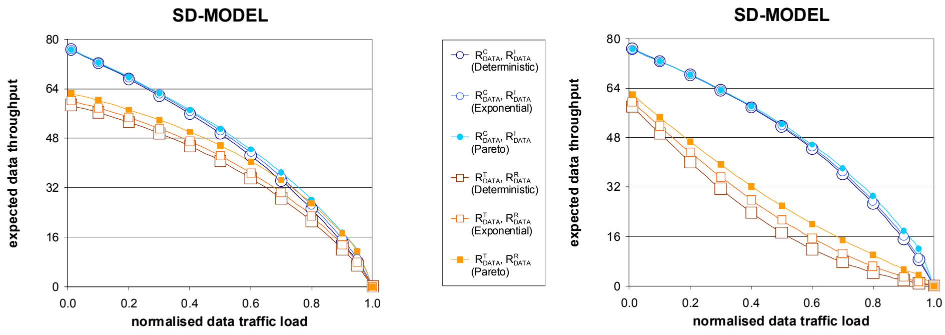


Figure 5.4 Comparison of different throughput measures in the SD model. All throughput measures are sensitive to the data call size distributions. The performance induced by three distinct distributions is shown.

from the perspective of the data traffic, the available capacity is deterministic at $C - \rho_{\text{speech}}\,(1 - \mathbf{P}_{\text{speech}})$, and hence the accelerated model corresponds with the D model. As a consequence, the associated curve is identical to the one for $\mathbf{R}^t_{\text{data}}$ in Figure 5.3 (right chart). Observe that as the capacity fluctuation process is accelerated, i.e. when ϑ is increased from 1 to ∞, the time-average throughput curves gradually approach the one corresponding to the extreme case of the D model, and the time-average throughput thus approximates the call-average throughput more and more closely. Additional numerical experiments (not included) indicate that among the different throughput measures, the ratio throughput measure is most sensitive to the degree of speech call dynamics in the SD model. While the call-average and expected instantaneous throughputs are largely insensitive to ϑ, and the time-average throughput converges to a significantly lower, yet positive value as $\vartheta \downarrow 0$, the ratio throughput measure becomes negligible for very small ϑ.

The right chart of Figure 5.5 shows the expected stretch of a data call for both the SD and D models. As noted in Section 5.4, the expected stretch in the D model is insensitive to the data call size distribution. For the SD model, such insensitivity does not hold, as is demonstrated by the three expected stretch curves for deterministic, exponential and Pareto (with shape parameter $\alpha = 1.35$) data call size distributions. Similar to the throughput performance, the expected stretch appears to be smaller (better) for more highly variable data call sizes. A noteworthy observation from the numerical experiments that is not included in the figure, is that the expected stretch

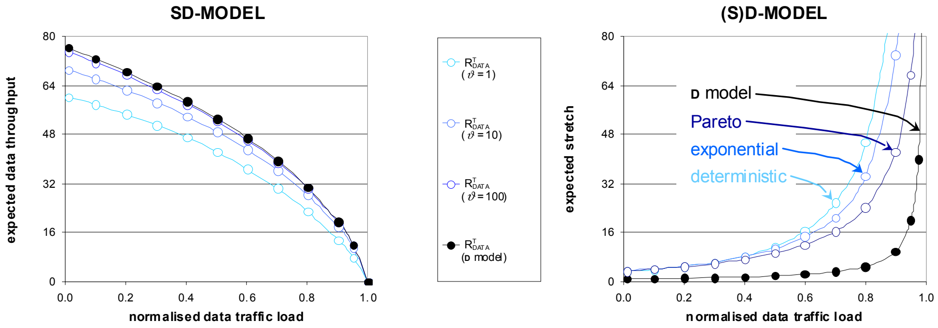


Figure 5.5 The impact of acceleration of the speech call arrival and departure process on $\mathbf{R}^t_{\text{data}}$ in the SD model (left chart). The expected stretch performance for different data call size distributions (SD model) as well as the insensitive values for the D model.

turns out to be infinitely large for the considered subexponential Weibull data call size distributions, i.e. with coefficient of variation greater than 1, for any data traffic load. In contrast, for highly variable Pareto distributions such as the one included in the figure, the expected stretch was finite within the stable regime of data traffic loads. The probable reason for this phenomenon is that a subexponential Weibull distribution features many very small data calls, which may suffer from excessively large relative sojourn times in the case of a varying service capacity that is even equal to zero at times. Pareto distributions are inherently truncated at the lower end, however, so that extremely small data calls do simply not occur. In any case, the expected stretch thus appears to be less useful as a measure for throughput performance.

5.6. CONCLUDING REMARKS

In this chapter we have specified, derived and compared, both analytically and numerically, a set of throughput measures in telecommunication systems serving elastic video or data calls according to a Processor Sharing service discipline. Although the investigations have been carried out for the specific context of a single cell in a (GSM/)GPRS network, the obtained insights are broadly applicable. The available capacity was either fixed, corresponding with a stand-alone dedicated GPRS network, or randomly varying, corresponding with an integrated services GSM/GPRS network, where the elastic calls utilise the capacity left idle by prioritised speech traffic. Among the considered throughput measures, the call-average throughput is considered to be the most appropriate indicator of the experienced Quality of Service. However, for models involving elastic calls of the data type, it is a hard measure to determine analytically. Among the alternative measures considered, the newly proposed and readily analytically derived *expected instantaneous throughput* measure is the only measure which excellently approximates (or is even equal to) the call-average throughput in all considered system models and across the entire range of considered elastic traffic loads. In particular for the practically most relevant model integrating speech and data traffic, other typically applied throughput measures such as the time-average throughput or the ratio of the expected call size and the expected sojourn time, significantly underestimate the call-average throughput. An intuitive reasoning for the generally (near-)perfect fit of the expected instantaneous throughput is that apparently, the throughput an elastic call experiences immediately upon arrival is an excellent predictor of what the call is likely to experience throughout its lifetime. Moreover, among the considered throughput measures, the expected instantaneous throughput is the *only* approximate measure that is truly *call*-centric.

The analytical evaluation further revealed that the expected call-average throughput of elastic video calls in the considered PS models is *insensitive* to both the variability of the available capacity and the call duration distribution, while the numerical experiments indicated that this insensitivity property also holds for the data service to a considerable degree. As seen in Chapter 4, the latter insensitivity does not hold if the data performance is measured by the (conditional) expected sojourn time.

CHAPTER 6

PERFORMANCE ANALYSIS OF DOWNLINK SHARED CHANNELS IN A UMTS NETWORK

WIRELESS data transfer is indisputably a major driver for the deployment and anticipated success of third-generation mobile networks such as the Universal Mobile Telecommunications System [45, 105]. In view of the commonly expected strong up/downlink data traffic asymmetry, the performance of the Downlink Shared transport CHannel (DSCH) is of specific interest among the different UMTS transport channels that have been standardised [1], as it most efficiently carries the anticipated heavy load of bursty downlink data traffic. As will be discussed below, the most relevant characteristics of the DSCHs are the channelisation code efficiency, and the requirement that each served data flow maintains an Associated Dedicated CHannel (A-DCH) in order to support fast closed-loop Transmission Power Control.

The experienced data performance on a DSCH is influenced by the data traffic load in two distinct manners. The most obvious *'round robin effect'* is that under a heavier data traffic load, the given DSCH channel rate is typically shared by a larger number of competing data calls so that an individual data call receives less attention from the 'server' and hence the experienced QOS is worse. On the other hand, the *'interference effect'* comprises of a reduced effective DSCH throughput as an increasing number of data calls and hence A-DCHs raises the interference levels and thus the BLock Error Rate, i.e. the probability that an arbitrary transmitted transport block is received erroneously: the greater the demand for service, the smaller the aggregate service capacity. The latter effect is further amplified in a multicellular scenario, where a DSCH experiences additional interference from the DSCHs and A-DCHs in surrounding cells, causing a further degradation of its effective throughput. The objective of the investigations in this chapter is to assess the DSCH performance in different scenarios with a gradually increasing degree of complexity, in order to identify the impact of the above-mentioned 'Round Robin' (RR) and 'interference effects'.

The outline of the chapter is as follows. Section 6.1 provides a brief review of the related literature, followed by a statement of this chapter's contribution in Section 6.2. A brief overview of the principal UMTS network aspects is given in Section 6.3. Section 6.4 then describes the model under investigation, while Section 6.5 presents the evaluation approach in terms of a decomposition of the general model into distinct experiments. The two-stage performance analysis of each experiment is outlined in Section 6.6. Subsequently, Section 6.7 presents a set of illustrative numerical experiments and discusses the results. Section 6.8 completes this chapter with some concluding remarks.

6.1. LITERATURE

A number of performance studies have been published that concentrate on performance aspects of CDMA-based networks. In line with the technology-driven system evolutions, only in the last few years researchers have shifted their focus from speech-only to integrated services CDMA networks, although the vast majority of the papers that concentrate on scheduling issues, consider packet scheduling in data-only scenarios. Given the commonly anticipated traffic asymmetry, virtually all recent data QOS investigations focus on downlink data transfer. As purely analytical studies in this field appear complicated and are therefore rare, most studies rely on dynamic simulations.

Comparisons of dedicated versus shared data transport channels are reported in [13, 22, 100, 116, 191]. [13] identifies optimality properties for downlink scheduling in CDMA-based data networks. It is proven that if the self-orthogonality factor, i.e. the loss of downlink orthogonality of differently delayed versions of the same signal, is smaller (better) than the cross-orthogonality factor, i.e. the orthogonality loss across distinct signals generated by the same NodeB, one-by-one scheduling outperforms simultaneous data transfers (on a per NodeB basis). This result advocates the use of shared (e.g. DSCHs) rather than dedicated channels for data transfer. Another interesting insight that is provided is that a higher resource utilisation can be achieved when remote users are scheduled in a one-by-one fashion, while simultaneous transfers are used to serve near users, in view of some (hardware-imposed) maximum that may hold on the per call data rate (hybrid scheme). In [22] an equivalent proof of the optimality of one-by-one scheduling over code-multiplexing (simultaneous transfers) as in [13] is given, which is subsequently exploited to conclude that one-by-one scheduling minimises total transfer time as well as the required energy. In view of the primary advantage of DSCH deployment of saving spreading codes and its main disadvantage of not allowing macro-diversity, [100] proposes and evaluates a hybrid scheme that serves remote data users on DCHs with the benefits of macro-diversity and near users on a DSCH. Note that this is the opposite of the hybrid scheme that was proposed in [13], which neglected the effects of macro-diversity. The hybrid scheme and proposed control policy for switching between DCHs and DSCHs were numerically demonstrated to achieve data rates as good as the pure DSCH scheme, but with considerably lower NodeB transmission power. The related uplink scheduling problem in a data-only CDMA-based network is studied in [116], using dynamic programming techniques. A numerical experiment is presented to compare the transfer time spans provided by simultaneous versus one-by-one scheduling depending on (exogenous) interference levels and data job sizes. Pure one-by-one scheduling is concluded to be optimal as long as the padding waste of incompletely filled transport blocks is negligible, e.g. in case data jobs are generally large. Closely related results are reported in [191], which analytically compares simultaneous transfers, one-by-one scheduling and a hybrid option for a system of delay-tolerant and delay-intolerant service calls sharing a CDMA-based uplink.

Different packet scheduling schemes are evaluated and compared in [7, 35, 107, 121, 137, 156]. In [7] an extensive simulation study of nine different rate- and delay-based scheduling algorithms is presented for the downlink of a single data-only CDMA cell, concentrating on the relevant trade-offs between efficiency and fairness. Another extensive simulation-based comparison of scheduling schemes is reported in [121]. Among the obtained results, it is concluded that schemes which exploit job sizes (assuming that such information is actually available) seem to outperform schemes that do not. Furthermore, whereas pure one-by-one scheduling (time-multiplexing) seems optimal when a continuous range of potential data rates is available, a combination of one-by-one and simultaneous transfers appears to be better if data rates are to be selected from a discrete set. In [137] the Rate Processor Sharing packet scheduling algorithm is presented and evaluated for the downlink of a single CDMA cell, which is equivalent to the widely adopted Discriminatory Processor Sharing scheme in wireline networking. As in the investigations presented in this chapter, a useful separation between the queueing and wireless aspects is made by characterising each user by a so-called effective weight, i.e. the amount of power required to support a unit data rate. Given these weights, queueing analysis is applied to determine the scheduling weights required to meet delay targets, and to determine rules for Call Admission Control. [35] concentrates on adaptive rate-controlled packet scheduling of data calls sharing a common transport channel in a single CDMA cell. The variations in the user-specific feasible data rates due to multipath propagation are exploited to enhance efficiency, while balancing the throughputs to provide fairness. The inherent trade-off between system throughput and user fairness is also considered in [107]. There a reference priority-based scheduling algorithm is studied, with the user priorities dynamically set to the ratio of the current C/I (carrier-to-interference ratio) and a window-averaged experienced throughput, which is noted to strike a good compromise between system throughput and fairness. In [33] multi-class processor sharing models are applied to assess user-level performance of channel-dependent scheduling in a single CDMA cell serving non-persistent data flows. The included numerical experiments indicate that a greedy and myopic scheduling algorithm (such as pure C/I-based scheduling) with maximises the instantaneous system throughput as well as the long-term average system throughput in a static scenario with persistent data flows, may yield suboptimal throughput performance in the more realistic dynamic scenario with non-persistent data flows. In [156] a Round Robin (fairness), C/I-based (high system throughput) and a hybrid packet scheduling scheme are compared via dynamic simulations, where the hybrid scheme is shown to be as fair as pure Round Robin while providing higher throughputs. As expected, pure C/I-based scheduling is demonstrated to provide a high system throughput but poor fairness. [88] presents a similar comparison study, including the effects of TCP flow control and also considers a mixed speech/data scenario. It was recently shown in , that Using analytical methods

It is noted that, with the exception of [156], none of the papers that either implicitly or explicitly concentrate on the DSCH performance, includes the potentially significant effects of the A-DCHs. Furthermore, some of the papers leave out the impact of intercellular interference caused by DSCH and A-DCH transmissions, focussing on single cell systems. The semi-analytical performance evaluation approach presented in this chapter is therefore applied to provide qualitative insight in the (relative) significance of these system aspects.

As a final reference, Chapter 7 assesses the performance gain that can be achieved by efficiently up- and downgrading of DSCH data rates in support of a varying presence of prioritised speech calls, using analysis in combination with Monte Carlo techniques. It is demonstrated that such adaptive scheduling can enhance both speech outage probabilities and data throughputs.

6.2. CONTRIBUTION

The presented study provides qualitative insight in the different system and traffic aspects that affect the DSCH performance in UMTS networks. In particular, we are interested in the specific impact of the 'Round Robin' and 'interference effects' on the data QOS under different data traffic loads. As the downlink orthogonality factor, i.e. the degree of non-orthogonality among signals generated by the same NodeB, influences the balance between intra- and intercellular interference, the sensitivity of the QOS with respect to this propagation environment-specific system parameter is also determined. A two-stage modelling approach is presented to segregate the interference aspects from the traffic dynamics, supporting an insightful analysis of the relevant system aspects.

6.3. UNIVERSAL MOBILE TELECOMMUNICATIONS SYSTEM

As a promising successor to the speech telephony-oriented second-generation systems such as the widely deployed Global System for Mobile communications [171], the third-generation Universal Mobile Telecommunications System [59, 105, 176] is specifically designed to efficiently support a wide variety of services with distinct Quality Of Service requirements. A brief overview of the UMTS network architecture, radio interface and transport channels is given below.

6.3.1. NETWORK ARCHITECTURE

Figure 6.1 depicts the most relevant components of the UMTS network architecture. As the current chapter concentrates on downlink data transfer, the figure includes an example flow between a data terminal and a remote server. Although the nomenclature differs, as well as the operational specifics of the network elements, the general structure of a UMTS network is very similar to that of a GPRS network (see Figure 3.1). In fact, the GSM/GPRS *core* network is reused, consisting of a circuit-switched segment of (G)MSCs connected to external PSTN/ISDN networks, and a packet-switched segment of S/GGSNs connected to external packet data networks. The entirely new UMTS terrestrial *radio access* network is composed of Radio Network Subsystems (RNSs), each consisting of a Radio Network Controller and a set of NodeBs that establish the physical link to the User Equipment. An important difference between the GSM/GPRS and

UMTS network architecture is that the RNCs, unlike GSM's BSCs, are connected directly in order to support soft handover across RNSs.

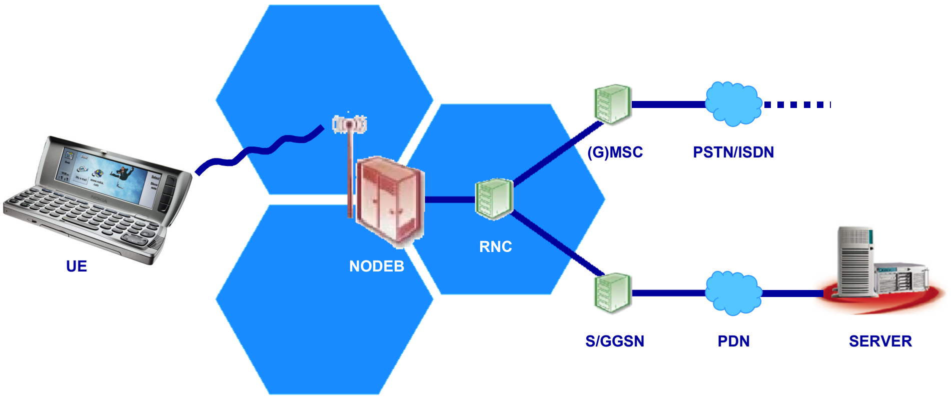


Figure 6.1 UMTS network architecture: the illustration shows an example data call maintained between the depicted UE and a remote server.

6.3.2. RADIO INTERFACE

Aside from the packet-oriented character also present in GPRS networks, the enhanced flexibility and resource efficiency of UMTS networks are in essence due to the newly standardised radio interface, which is based on the Code Division Multiple Access [218] radio transmission technology. The relatively wide carrier bandwidth of 5 MHz in combination with the use of Orthogonal Variable Spreading Factor (OVSF) channelisation codes (see e.g. [105, 138]) enables the flexible assignment of a large range of bit rates (up to 2 Mbits/s in a low/no mobility scenario) to a potentially broad range of services. The enhanced resource efficiency is primarily due to the CDMA-inherent phenomenon that any radio resources left idle by a given call or in a given cell, are implicitly available to other calls/cells, which is a direct consequence of the applied universal frequency reuse and the fact that CDMA network capacity is determined by the experienced interference levels. This inherent flexibility is in stark contrast to GSM networks, where Dynamic Channel Allocation (DCA) schemes have been specified to pursue the explicit spatial reallocation of idle capacity, but both the involved granularity of capacity and the inherent frequency reuse restrictions severely complicate matters, to the extent that few GSM network operators deploy DCA.

6.3.3. TRANSPORT CHANNELS

As the foreseen services greatly differ in their traffic characteristics and QOS requirements, a number of distinct transport channels have been specified to accommodate these services most efficiently [1, 105]. Specifically designed for delay-sensitive services or services with stringent throughput requirements, a D*edicated* transport CHannel is a 'bit pipe' associated with a single UE and has the advantage of fast closed-loop Transmission Power Control and macro-diversity. On the other hand, a set of *common* transport channels is specified, which may be shared by multiple UEs. Among these, the Random and Forward Access CHannels (RACH/FACH) are typically used for the transfer of relatively small data chunks in up- and downlink, respectively, without the advantages of closed-loop Transmission Power Control and macro-diversity. Medium- to large data transfers, particularly of bursty character (e.g. TCP/IP flows), are most efficiently conveyed on the uplink Common Packet CHannel (CPCH) or the Downlink Shared CHannel (DSCH). The CPCH and the DSCH enjoy the advantages of closed-loop Transmission Power Control by maintaining an Associated DCH (A-DCH) for each call to carry control signalling information, but the use of macro-diversity is rather complicated from an implementation viewpoint and therefore not standardised. The principal advantage of the DSCH is the enhanced efficiency of OVSF code usage: since downlink *bursty* data flows are anticipated to make up the bulk of the traffic, and many such relatively low-activity flows can be concurrently active, time-sharing a common code (DSCH) places a significantly milder claim on the OVSF code tree than assigning each such call its own code. Note that the low bit rate A-DCHs associated with each data flow that is handled on a DSCH, require only one of the many available long OVSF codes.

6.4. MODEL

This section sets the framework for the presented performance analysis by describing the system, propagation and traffic models in generic terms. Concrete parameter settings are specified in Table 6.1 at the beginning of Section 6.7 below.

6.4.1. SYSTEM MODEL

Consider a UMTS network of B hexagonal cells with a given cell radius served by omnidirectional NodeBs (see Figure 6.2). In view of the anticipated dominance of wireless data services and the expected up-/downlink data traffic load asymmetry, the presented investigation concentrates on the modelling and performance evaluation of UMTS's DSCH. The system model assumes the deployment of a single DSCH with a given nominal bit rate R_{DSCH} at each NodeB $b \in \mathbb{B} \equiv \{1, 2, \cdots, B\}$ and the energy-per-bit to interference-plus-noise density ratio (E_b/N_o) target γ_{DSCH} (the physical layer radio link quality requirement) which multiplexes the present data calls according to an idealised Round Robin scheduling discipline.

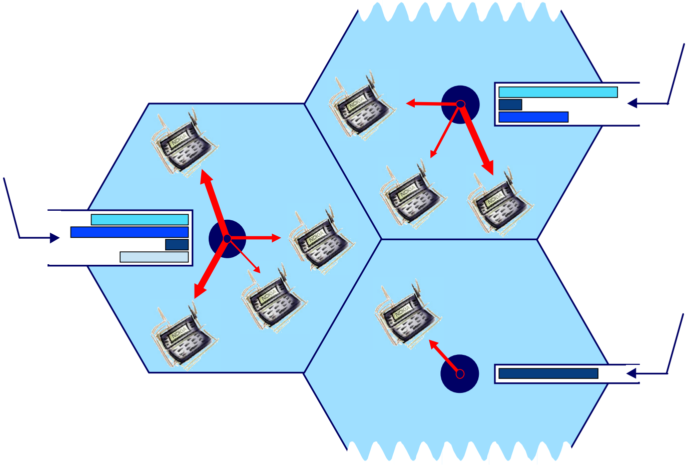


Figure 6.2 Illustration of the system model, i.e. a UMTS network of B (h.l. $B = 3$) hexagonal cells served by omnidirectional NodeBs. Each NodeB features a single DSCH to multiplex the different data flows.

Each data call that is served on the DSCH maintains a low bit rate A-DCH for control signalling purposes, a.o. to support closed-loop Transmission Power Control. The A-DCHs are characterised by bit rate $R_{\text{A-DCH}}$ and E_b/N_o target $\gamma_{\text{A-DCH}}$. The closed-loop Transmission Power Control maintained on the A-DCH enables power efficient DSCH transmissions, by applying a fixed DSCH/A-DCH power offset of

$$\frac{p_{\text{DSCH}}}{p_{\text{A-DCH}}} = \frac{\widetilde{\gamma}_{\text{DSCH}}}{\widetilde{\gamma}_{\text{A-DCH}}} \Leftrightarrow p_{\text{DSCH}} = p_{\text{A-DCH}} \left(\frac{\gamma_{\text{DSCH}} R_{\text{DSCH}}}{\gamma_{\text{A-DCH}} R_{\text{A-DCH}}} \right), \tag{6.1}$$

where R_c denotes the system chip rate, $\widetilde{\gamma}_{\text{DSCH}} \equiv \gamma_{\text{DSCH}} (R_{\text{DSCH}}/R_c)$ is the DSCH's C/I target, and $\widetilde{\gamma}_{\text{A-DCH}} \equiv \gamma_{\text{A-DCH}} (R_{\text{A-DCH}}/R_c)$ is the A-DCH's C/I target (the ratio of the system chip rate and information bit rate, e.g. R_c/R_{DSCH}, is generally referred to as the *processing gain*). Such a fixed power offset can be applied because the DSCH and the A-DCH transmissions share the same propagation paths from the serving NodeB to the UE, and hence experience identical path gain variations. Although in general DCHs allow the use of soft handover to enhance radio link quality and resource efficiency, this feature does not apply to the A-DCHs as it would preclude the proper use of a fixed DSCH/A-DCH power offset for DSCH Transmission Power Control (recall that the DSCH does not support soft handover). The NodeB's transmission power budget is denoted $p_{\max}$, and the spatially uniform thermal noise level is given by $\nu > 0$.

6.4.2. PROPAGATION MODEL

The radio propagation model considers a single signal path with correlated lognormal shadowing and Rayleigh fading (see also Section 1.2.3). Given a distance r between the transmitter NodeB and the receiver UE, the relation between transmission ($p_{\text{transmission}}$) and the instantaneous reception ($p_{\text{reception}}$) power (in Watt) is given by

$$\begin{aligned} p_{\text{reception}} &= p_{\text{transmission}} \cdot \mathcal{G}_{\text{UE NodeB}} \\ &= p_{\text{transmission}} \cdot \eta_{\text{basic}} \cdot r^{-\varsigma} \cdot 10^{(a\xi_{\text{UE}}+b\xi_{\text{NodeB}})/10} \cdot \zeta_{\text{Rayleigh}}, \end{aligned}$$

where η_{basic} reflects the basic transmission loss, ς is the attenuation exponent, $a\xi_{\text{UE}} + b\xi_{\text{NodeB}}$ is the correlated shadowing effect (in dB), with ξ_{UE}, $\xi_{\text{NodeB}} \sim N\left(0, \sigma_{\text{S}}^2\right)$ the mutually independent UE- and NodeB-specific shadowing effects and a and b the correlation factors [218], and $\zeta_{\text{Rayleigh}} \sim Exp\,(1)$ the random (instantaneous) Rayleigh fading effect (e.g. [141]). The slow and fast fading effects are also assumed to be independent. The *local average* path gain, i.e. excluding the Rayleigh fading effect and denoted by $\mathcal{G}^{\bullet}_{\text{UE NodeB}}$, is used to assign a serving NodeB to a given UE. In C/I calculations, ω denotes the downlink orthogonality factor, reflecting the degree of orthogonality of signals generated by the same NodeB, assumed to be the same for time-shifted versions of identical or different signals from the same NodeB. The value of ω typically depends on the propagation environment.

6.4.3. TRAFFIC MODEL

The considered UMTS network serves data calls only, assumed to be downlink transfers of files with exponentially distributed sizes. The mean file size is denoted by μ^{-1} (in kbits). File transfer requests are generated according to a Poisson process with rate λ_b in cell $b \in \mathbb{B}$, and terminals are uniformly distributed over the cell of origination. Note that a data call originating in cell $b \in \mathbb{B}$ (i.e. geographically nearest to NodeB b) may be nearer (in the *path gain* sense) to NodeB $b' \neq b$ due to the shadowing effects. Without loss of generality this effect is assumed to be captured by the λ_b's, so that λ_b is the average arrival rate of data calls that are nearest (in the path gain sense) to NodeB $b \in \mathbb{B}$. Let $\rho_b \equiv \lambda_b / (\mu R_{\text{DSCH}})$ denote the (normalised) data traffic load in cell $b \in \mathbb{B}$. Two distinct Call Admission Control thresholds are enforced. Firstly, a UE whose location (read: local average path gain) is so unfortunate that its C/I requirement cannot be met even in an otherwise empty system, i.e.

$$\max_{b \in \mathbb{B}} \mathcal{G}^{\bullet}_{\text{UE}\ b} < \frac{\nu}{\left(\dfrac{R_c}{\gamma_{\text{DSCH}} R_{\text{DSCH}} + \gamma_{\text{A-DCH}} R_{\text{A-DCH}}} - \omega \right) p_{\max}}, \tag{6.2}$$

is rejected in order to avoid serving 'hopeless' calls. In the above path gain threshold, which is derived using (6.1), the term '$\gamma_{\text{A-DCH}} R_{\text{A-DCH}}$' should be omitted if no A-DCHs are considered. The second CAC threshold limits the number of data calls in service to $a_{\max}$ in each cell in order to provide some minimum QOS, so that the space of feasible system states is given by

$$\mathbb{S} \equiv \left\{ \mathbf{a} \equiv (a_1, a_2, \cdots, a_B) \in \{0, \cdots, a_{\max}\}^B \right\},$$

where a_b denotes the number of data calls in cell $b \in \mathbb{B}$. Let $\mathbb{S}_b^+ \equiv \{\mathbf{a} \in \mathbb{S} : a_b > 0\}$ denote the set of states in which at least one data call is served at NodeB $b \in \mathbb{B}$. The setting of the CAC threshold $a_{\max}$ is addressed in Section 6.6 and demonstrated in Section 6.7 below. In principle, the presented analysis allows the CAC decision to be less myopic and also includes the actual loading at e.g. adjacent cells, but for our qualitative purposes we chose to keep the CAC scheme simple.

6.4.4. PERFORMANCE MEASURES

The performance of the data calls is expressed in terms of the *expected transfer time*. Additionally, the *conditional expected transfer time* of a data call of a given size is derived. As an intermediate, physical layer performance measure, the *outage* *probability* is applied, defined as the likelihood that a radio link does not meet its carrier-to-interference ratio requirement at an arbitrary transmission attempt.

6.5. EVALUATION APPROACH

In order to obtain the intended qualitative insight regarding the 'Round Robin effect' and the 'interference effects' of the traffic load on the DSCH performance, the general framework described in the previous section is decomposed into five distinct experiments, as visualised in Figure 6.3. The most basic EXPERIMENT 0 considers a single NodeB (and DSCH) without any interference or other radio interface-specific effects and thus captures the 'Round Robin effect'. The different 'interference effects' are captured by the subsequent experiments. EXPERIMENT 1A adds the DSCH's self-interference due to orthogonality loss, while EXPERIMENT 1B further adds the interference from the maintained A-DCHs, still considering a single NodeB. The impact of the inter-NodeB interference from the DSCH and A-DCHs is investigated by adding additional NodeBs in EXPERIMENTS 2 and 3. Although the evaluation method readily extends to an arbitrary number of NodeBs, the largest network size is set to three NodeBs in order to allow reasonably swift analytical calculations.

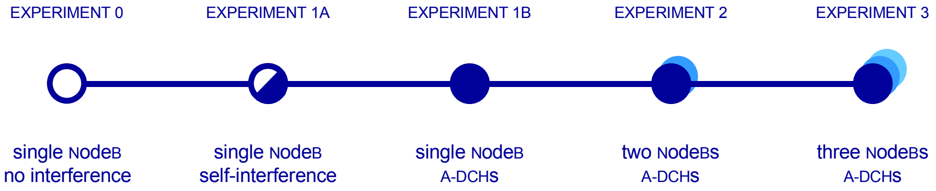


Figure 6.3 The evaluation approach consists of a sequence of experiments that concentrate on gradually more involved scenarios in order to assess the relevance of different system aspects.

6.6. PERFORMANCE ANALYSIS

The principal objective of this section is to outline the performance analysis required for EXPERIMENTS 2 and 3, including all relevant system aspects, while the specific characteristics of the simpler EXPERIMENTS 0 and 1 are discussed in a separate paragraph at the end of each subsection.

As in a CDMA-based network, the precise locations of the served calls have a significant impact on the required transmission powers, the corresponding interference levels, the induced BLock Error Rates and hence the experienced throughputs, the performance evaluation approach generally encountered in similar models, is that of (time-consuming) dynamic system-level simulations. Given our qualitative objective to generate insight into the DSCH performance and its sensitivity with respect to e.g. the presence of the A-DCHs, the data traffic load and the downlink orthogonality factor, we propose a hybrid approach in two stages, which segregates the interference aspects from the traffic dynamics (along similar lines as followed in Chapter 8, as well as in [79, 137]).

In STAGE I a Monte Carlo simulation experiment is carried out to determine the outage probability that a reference call in cell $b \in \mathbb{B}$ experiences during its DSCH transfer as a function of the number of calls in each cell and *conditional* on its admission by the path gain-based CAC criterion (6.2). The set of outage probability functions is denoted $\left\{P_b\left(\mathbf{a}\right),\ \mathbf{a} \in \mathbb{S}_b^+,\ b \in \mathbb{B}\right\}$, and captures the random effects of terminal location and signal fading. The Monte Carlo simulations are also readily utilised to determine the fraction $\mathbf{P}_f^{\mathcal{G}}$ of calls that fail to satisfy the minimum path gain requirement and are thus rejected.

STAGE II captures the traffic dynamics of call arrivals and terminations in a B-dimensional irreducible continuous-time Markov chain $\left(\mathbf{A}(t)\right)_{t\geq 0} \equiv (A_1(t), A_2(t), \cdots, A_B(t))_{t\geq 0}$ with system states $\mathbf{a} \in \mathbb{S}$, specified by the effective cell-specific call arrival rates $\lambda_b\left(1 - \mathbf{P}_f^{\mathcal{G}}\right)$ for a given gross fresh call arrival rate λ_b, $b \in \mathbb{B}$, the CAC threshold $a_{\max}$, and the expected effective throughput per data call given by

$$\beta_b\left(\mathbf{a}\right) \equiv a_b^{-1}\left(1 - P_b\left(\mathbf{a}\right)\right) R_{\text{DSCH}},$$

as experienced in cell b in system state $\mathbf{a} \in \mathbb{S}$. Note that we may indeed apply the reduced arrival process with rate $\lambda_b\left(1 - \mathbf{P}_f^{\mathcal{G}}\right)$ to the considered Markov chain, as it is still Poisson due to the random filtering of badly located calls. Data scheduling is assumed to be done according to the Processor Sharing discipline, which is an idealised and analytically more tractable version of the Round Robin discipline. Here the outage probability functions obtained in STAGE I are interpreted as the experienced BLERs (see also Remark 6.2 below). The data QOS can be derived from the equilibrium and transient behaviour of the Markov chain.

Both stages are treated in more detail below, starting with STAGE I on interference aspects.

6.6.1. INTERFERENCE ASPECTS

In STAGE I the set of functions $\left\{P_b(\mathbf{a}),\ \mathbf{a}\in\mathbb{S}_b^+,\ b\in\mathbb{B}\right\}$ is determined by means of Monte Carlo simulations, including the derivation of the CAC threshold $a_{\max}$ which bounds $\mathbb{S}_b$, $b\in\mathbb{B}$. For each feasible system state $\mathbf{a}\in\mathbb{S}_b^+$, K independent constellations (or snapshots) $(\mathbb{A},\mathbb{D},\mathcal{G},\mathbf{b})^k$ are generated, $k=1,\cdots,K$ (for enhanced readability, the index k indicating the considered constellation is omitted throughout the section). Let $A\equiv\left(\sum_{b\in\mathbb{B}}a_b\right)$ denote the total number of data calls in the network in a given state $\mathbf{a}\in\mathbb{S}_b^+$. A constellation is specified by the set of all data calls $\mathbb{A}\equiv\{1,\cdots,A\}$, the subset $\mathbb{D}\subset\mathbb{A}$ of data calls that are (randomly) selected (one per non-idle NodeB) for service by the Round Robin scheduler in the considered constellation, the $(A\times B)$-dimensional path gain matrix $\mathcal{G}$, and the A-dimensional base station assignment vector $\mathbf{b}\equiv(b_1,\cdots,b_A)$, where b_m denotes the NodeB serving data call $m\in\mathbb{A}$, with $|\{m\in\mathbb{A}:b_m=b\}|=a_b$, $b\in\mathbb{B}$. In view of the path gain-based CAC criterion the Monte Carlo experiments are designed to ensure that the sampled data calls satisfy the minimum dominant path gain requirement (6.2).

In order to determine whether any of the active data calls experiences an outage in a given constellation $(\mathbb{A},\mathbb{D},\mathcal{G},\mathbf{b})$, it is necessary to verify whether a vector $\mathbf{p}\equiv\left((\widetilde{p}_m,m\in\mathbb{A}),\ (\widehat{p}_m,m\in\mathbb{D})\right)$ of A-DCH and DSCH transmission powers exists, respectively, which satisfies the signals' C/I requirements:

$$\begin{cases} \dfrac{\widetilde{p}_m\mathcal{G}_{mb_m}}{I_m(\mathbf{p})+\nu}\geq\widetilde{\gamma}_{\text{A-DCH}} & m\in\mathbb{A},\\[2ex] \dfrac{\widehat{p}_m\mathcal{G}_{mb_m}}{I_m(\mathbf{p})+\nu}\geq\widetilde{\gamma}_{\text{DSCH}} & m\in\mathbb{D},\\[2ex] \displaystyle\sum_{m\in\mathbb{A}:b_m=b}\widetilde{p}_m+\sum_{m\in\mathbb{D}:b_m=b}\widehat{p}_m\leq p_{\max} & b\in\mathbb{B}, \end{cases}$$

where the different sets of conditions reflect the C/I requirements on the A-DCHs and DSCHs, and the power budget of the NodeBs, and

$$I_m(\mathbf{p}) \equiv \sum_{m' \in \mathbb{A}} \omega(b_m, b_{m'})\, \widetilde{p}_{m'} \mathcal{G}_{mb_{m'}} + \sum_{m' \in \mathbb{D}} \omega(b_m, b_{m'})\, \widehat{p}_{m'} \mathcal{G}_{mb_{m'}},$$

denotes the total interference power experienced by data call $m \in \mathbb{A} \cup \mathbb{D}$, given transmission power vector $\mathbf{p}$, and

$$\omega(b, b') \equiv \begin{cases} \omega & \text{if } b = b' \\ \\ 1 & \text{if } b \neq b' \end{cases}$$

gives the appropriate downlink orthogonality factor.

The section below elaborates on the determination of the optimal power vector that satisfies the above system of inequalities if the constellation is feasible, or the suboptimal power vector in case it is infeasible and all C/I's in the congested cell(s) are proportionally downgraded to meet the power budget restriction(s). Whether or not any of the DSCHs experiences an outage can be determined immediately from the achieved (sub)optimal power vector. It is noted that even if the system as a whole is not feasible, possibly due to one congested cell, it is possible that the DSCH C/I target is satisfied in some, more lightly loaded, cells sufficiently far from the bottleneck cell(s).

INTERMEZZO: ON CONSTELLATION (IN)FEASIBILITY

Consider a constellation $(\mathbb{A}, \mathbb{D}, \mathcal{G}, \mathbf{b})$ where $\mathbb{A}$ denotes the set of A-DCHs, $\mathbb{D}$ denotes the active DSCH links, matrix $\mathcal{G}$ contains the path gains between terminals and NodeBs, and $\mathbf{b}$ is the NodeB assignment vector. Let C/I targets associated with the A-DCHs and DSCHs be denoted $\widetilde{\gamma}_{\text{A-DCH}}$ and $\widetilde{\gamma}_{\text{DSCH}}$, respectively. Using straightforward algebraic manipulations, the C/I conditions for A-DCH $m \in \mathbb{A}$ can be written as

$$\widetilde{p}_m - \sum_{\substack{m' \in \mathbb{A} \\ m' \neq m}} \widetilde{p}_{m'} \frac{\omega(b_m, b_{m'})\, \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} - \sum_{m' \in \mathbb{D}} \widehat{p}_{m'} \frac{\omega(b_m, b_{m'})\, \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \geq \frac{\nu}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}},$$

while analogously, the C/I condition for DSCH $m \in \mathbb{D}$ is given by

$$\widehat{p}_m - \sum_{m' \in \mathbb{A}} \widetilde{p}_{m'} \frac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{DSCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} - \sum_{\substack{m' \in \mathbb{D} \\ m' \neq m}} \widehat{p}_{m'} \frac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{DSCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \geq \frac{\nu}{\left(\widetilde{\gamma}_{\text{DSCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}}.$$

Obvious necessary conditions for feasibility are $\omega\widetilde{\gamma}_{\text{A-DCH}} < 1$ and $\omega\widetilde{\gamma}_{\text{DSCH}} < 1$. The orthogonality function $\omega\left(b, b'\right)$ is defined in Section 6.6.1. The C/I conditions of all A-DCHs and DSCHs are readily collected in the matrix form given by

$$\left(\mathcal{I} - \mathcal{H}\right) \mathbf{p} \geq \widehat{\boldsymbol{\nu}}, \tag{6.3}$$

where $\mathbf{p} \equiv \left(\widetilde{p}_m,\ m \in \mathbb{A}; \widehat{p}_m,\ m \in \mathbb{D}\right)$, the $\left(|\mathbb{A}| + |\mathbb{D}|\right) \times \left(|\mathbb{A}| + |\mathbb{D}|\right)$-dimensional matrix $\mathcal{H}$ is partitioned as follows:

$$\begin{aligned} \mathcal{H} &\equiv \begin{pmatrix} \mathcal{H}_{\mathbb{A}\mathbb{A}} & \mathcal{H}_{\mathbb{A}\mathbb{D}} \\ \mathcal{H}_{\mathbb{D}\mathbb{A}} & \mathcal{H}_{\mathbb{D}\mathbb{D}} \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} 1\left\{m \neq m'\right\} & \dfrac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \\ \\ \dfrac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{DSCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} & \dfrac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{DSCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} 1\left\{m \neq m'\right\} \end{pmatrix}, \end{aligned}$$

where $m, m' \in \mathbb{A}$ in $\mathcal{H}_{\mathbb{A}\mathbb{A}}$, $m \in \mathbb{A}$ and $m' \in \mathbb{D}$ in $\mathcal{H}_{\mathbb{A}\mathbb{D}}$, $m \in \mathbb{D}$ and $m' \in \mathbb{A}$ in $\mathcal{H}_{\mathbb{D}\mathbb{A}}$, and $m, m' \in \mathbb{D}$ in $\mathcal{H}_{\mathbb{D}\mathbb{D}}$. The $\left(|\mathbb{A}| + |\mathbb{D}|\right)$-dimensional vector $\widehat{\boldsymbol{\nu}}$ is given by

$$\begin{aligned} \widehat{\boldsymbol{\nu}} &\equiv \left(\widehat{\boldsymbol{\nu}}_{\mathbb{A}}, \widehat{\boldsymbol{\nu}}_{\mathbb{D}}\right) \\ &= \left(\left(\frac{\nu}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}},\ m \in \mathbb{A}\right); \left(\frac{\nu}{\left(\widetilde{\gamma}_{\text{DSCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}},\ m \in \mathbb{D}\right)\right). \end{aligned}$$

Momentarily disregarding the power budget constraint per NodeB, given by

$$\sum_{\substack{m \in \mathbb{A} \\ b_m = b}} \widetilde{p}_m + \sum_{\substack{m \in \mathbb{D} \\ b_m = b}} \widehat{p}_m \leq p_{\max}, \tag{6.4}$$

$b \in \mathbb{B}$, it can be shown that a power assignment $\mathbf{p} > 0$ that satisfies (6.3) exists if and only if $(\mathcal{I} - \mathcal{H})^{-1}$ exists and is positive component-wise [12, 85, 170]. Also, if such a $\mathbf{p}$ exists, then

$$\mathbf{p}^{\star} \equiv (\mathcal{I} - \mathcal{H})^{-1}\,\widehat{\boldsymbol{\nu}} = \sum_{j=0}^{\infty} \mathcal{H}^{j}\widehat{\boldsymbol{\nu}} \tag{6.5}$$

is a Pareto optimal solution to the power assignment problem, in the sense that any other feasible power assignment $\mathbf{p}$ would require at least as high a transmission power for each UE, i.e. $p_m \geq p_m^{\star}$ for all $m \in \mathbb{A} \cup \mathbb{D}$. Hence *if* the constellation $(\mathbb{A}, \mathbb{D}, \mathcal{G}, \mathbf{b})$ is feasible then the C/I requirements can be satisfied, *with equality*, simultaneously on all radio links.

Remark 6.1 We adopt a slightly loose notion of Pareto optimality here. Strictly speaking, a given power assignment is Pareto optimal if no other power assignment exists for which all radio links are as good while at least one is strictly better (e.g. [135]). The looseness of the definition as applied in the considered context lies therein that the posed objective is not to *optimise* the radio link quality, but to ensure that all radio links are *sufficiently good*, while minimising the required transmission powers. Although this is the power control objective as applied in live networks, we note that in many constellations, it is in principle possible to assign 'excessive' transmission powers in order to achieve an above-target C/I, a consequently below-target BLER and thus a higher effective data throughput (see also Remark 6.2).

Clearly, a constellation which is infeasible when disregarding power budget constraint (6.4) posed by $p_{\max}$, i.e. assuming $p_{\max} = \infty$, is also infeasible under the additional constraint posed by the NodeBs' power budgets. Furthermore, given a constellation which is feasible for $p_{\max} = \infty$, it is also feasible for $p_{\max} < \infty$ if and only if $\mathbf{p}^{\star}$ as given by (6.5) satisfies the power budget conditions (6.4), due to the Pareto optimality of $\mathbf{p}^{\star}$.

For a feasible constellation, the final expression in (6.5) suggests the following iterative Transmission Power Control procedure

$$\mathbf{p}(j+1) = \mathcal{H}\,\mathbf{p}(j) + \widehat{\boldsymbol{\nu}}, \tag{6.6}$$

initialised by

$$\mathbf{p}(0) = \widehat{\boldsymbol{\nu}},$$

which has a logical interpretative meaning, as it constitutes optimal power settings in the presence of thermal noise only, e.g. for $m \in \mathbb{A}$:

$$\widetilde{p}_m(0) = \widehat{\nu} = \frac{\nu}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \Leftrightarrow \frac{\widetilde{p}_m(0)\mathcal{G}_{mb_m}}{\omega \widetilde{p}_m(0)\mathcal{G}_{mb_m} + \nu} = \widetilde{\gamma}_{\text{A-DCH}}.$$

Note that the Transmission Power Control update step also has a logical interpretative meaning, e.g. for $m \in \mathbb{A}$:

$$\begin{aligned} \widetilde{p}_m(j+1) &= \sum_{\substack{m' \in \mathbb{A} \\ m' \neq m}} \left(\frac{\omega(b_m, b_{m'}) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \right) \widetilde{p}_{m'}(j) \\ &\quad + \sum_{m' \in \mathbb{D}} \left(\frac{\omega(b_m, b_{m'}) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \right) \widehat{p}_{m'}(j) + \frac{\nu}{\left(\widetilde{\gamma}_{\text{A-DCH}}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \\ &\Leftrightarrow \frac{\widetilde{p}_m(j+1)\mathcal{G}_{mb_m}}{\omega \widetilde{p}_m(j+1)\mathcal{G}_{mb_m} + \widehat{I}_m(\mathbf{p}(j)) + \nu} = \widetilde{\gamma}_{\text{A-DCH}}, \end{aligned}$$

where

$$\widehat{I}_m(\mathbf{p}(j)) \equiv \sum_{\substack{m' \in \mathbb{A} \\ m' \neq m}} \omega(b_m, b_{m'}) \widetilde{p}_{m'}(j) \mathcal{G}_{mb_{m'}} + \sum_{m' \in \mathbb{D}} \omega(b_m, b_{m'}) \widehat{p}_{m'}(j) \mathcal{G}_{mb_{m'}}$$

denotes the total interference power experienced by data call m from all *other* signals, given the transmission power vector $\mathbf{p}(j)$ determined in the previous TPC iteration.

Hence A-DCH m's transmission power is updated such that its C/I calculated from the other links' *previous* transmission powers and its own *new* power, precisely equals the target value $\widetilde{\gamma}_{\text{A-DCH}}$. This implies that the proposed Transmission Power Control scheme can be implemented in a distributed manner. Since $\mathcal{H}$ has non-negative elements only, the transmission powers increase during the proposed distributed Transmission Power Control procedure.

In accordance with [12], we note that the Transmission Power Control update procedure (6.6) can be simplified to

$$\widetilde{p}_m(j+1) = \widetilde{p}_m(j) \cdot \left\{ \frac{(C/I)^{-1}_{m,j} - \omega}{\widetilde{\gamma}^{-1}_{\text{A-DCH}} - \omega} \right\}, \tag{6.7}$$

$m \in \mathbb{A}$, where $(C/I)_{m,j}$ denotes A-DCH m's experienced carrier-to-interference ratio after the jth Transmission Power Control iteration ([12] assumes $\omega = 0$ which further simplifies the power update factor to the ratio of the target and current C/I). An analogous simplification is readily given for the DSCHs. If the constellation $(\mathbb{A}, \mathbb{D}, \mathcal{G}, \mathbf{b})$ is feasible the basic Transmission Power Control procedure given by (6.6) or its simplified equivalent (6.7) can be shown to converge at a geometric rate (see e.g. [22]).

In order to deal with the possible infeasibility of a constellation $(\mathbb{A}, \mathbb{D}, \mathcal{G}, \mathbf{b})$ (under a finite $p_{\max}$), each Transmission Power Control iteration step is followed by the adjustment step

$$\widetilde{p}'_m(j+1) = \widetilde{p}_m(j+1) \cdot \min \left\{ 1, \frac{p_{\max}}{\sum\limits_{\substack{m' \in \mathbb{A} \\ b_{m'} = b_m}} \widetilde{p}_{m'}(j+1) + \sum\limits_{\substack{m' \in \mathbb{D} \\ b_{m'} = b_m}} \widehat{p}_{m'}(j+1)} \right\},$$

$m \in \mathbb{A}$ (and analogously for the DSCHs), to avoid power divergence and ensure that the aggregate transmission power per NodeB remains within its budget. Note that the transmission *power* of each call in a congested cell is reduced proportionally, which corresponds with an approximately proportional reduction of the newly experienced C/I's, given a typically dominant *inter*cellular interference. It is stressed that in an infeasible constellation, it is not necessarily the case that the established C/I falls below the target value on *all* radio links.

PERFORMANCE MEASURES

If for a given system state $\mathbf{a} \in \mathbb{S}_b^+$, NodeB b's DSCH experienced an outage in $o_b(\mathbf{a})$ out of K snapshots, the outage probability is estimated at

$$P_b(\mathbf{a}) \equiv \frac{o_b(\mathbf{a})}{K}, \; b \in \mathbb{B}.$$

The outage probability functions $\left\{P_b\left(\mathbf{a}\right),\ \mathbf{a} \in \mathbb{S}_b^+,\ b \in \mathbb{B}\right\}$ are increasing in the DSCH transfer rate R_{DSCH}, due to the more demanding C/I requirements, and decreasing in ω, due to the reduced experienced interference. As the $P_b\left(\mathbf{a}\right)$'s are specified for a given realisation of the system state, they do not depend on the traffic dynamics. Note that the obtained set of functions $\left\{P_b\left(\mathbf{a}\right),\ \mathbf{a} \in \mathbb{S}_b^+,\ b \in \mathbb{B}\right\}$, with $P_b\left(\mathbf{a}\right)$ increasing in each a_b, $b \in \mathbb{B}$, implicitly indicates the amount of resources that a data call at a given NodeB claims at all other NodeBs in terms of a reduced effective DSCH throughput, and is in that sense similar to the effective interference models that have been derived for DCHs in CDMA-based networks (see e.g. Chapter 8 and [79]). The CAC threshold $a_{\max}$ is derived from the obtained outage probability functions in combination with an operator's outage (or BLER) target value, thereby effectively bounding $\mathbb{S}_b$ (and hence $\mathbb{S}_b^+$), for each $b \in \mathbb{B}$.

Remark 6.2 As stated before, in STAGE II of the presented analysis the obtained outage probabilities will be interpreted as BLock Error Rates. This approximation effectively assumes that an established C/I which meets its target value invariably leads to a successfully decoded transport block, while an insufficient C/I consistently implies an erroneous transport block. In practice, the relation between the experienced C/I and the corresponding BLER follows a more gradual continuous curve that is generally obtained by means of detailed link level simulations, which include all relevant physical layer effects such as channel coding, interleaving, modulation and multipath fading. The discontinuous relation assumed here for our qualitative purposes *over*estimates the BLER for insufficient C/I's, yet *under*estimates the BLER for C/I's that do meet the target level, while the net effect of the approximation naturally depends on the C/I distribution. We note that in case an applicable C/I-versus-BLER curve is at one's disposal, the STAGE I experiments are readily adjusted to deliver a true BLER function for each feasible system state, rather than the outage probability function.

OTHER EXPERIMENTS

The STAGE I Monte Carlo analyses required to obtain the outage probabilities of the simpler experiments are considerably less extensive than those needed for the most complete EXPERIMENT 3. Recall that the basic reference EXPERIMENT 0 excludes all radio interface-related aspects and therefore requires no results from STAGE I ($\mathbf{P}_f^{\mathcal{G}} = P_1\left(a_1\right) = 0$, $a_1 \in \mathbb{S}_1^+$). The outage probability in EXPERIMENT 1A is influenced merely by whether the considered DSCH is on or off, and not by the *number* of present data calls, due to the absence of A-DCHs. Therefore only a single outage probability $P_1^{\bullet}(1)$ is required, while the A-DCH related C/I requirements are redundant in determining whether the reference DSCH experiences an outage in a given constellation. The analysis for EXPERIMENT 1B is equivalent to that for EXPERIMENTS 2 and 3, requiring $a_{\max}$ outage probabilities $\left\{P_1(a_1),\ a_1 \in \mathbb{S}_1^+\right\}$.

6.6.2. TRAFFIC DYNAMICS

Given the outage probability functions $\left\{P_b(\mathbf{a}),\ \mathbf{a} \in \mathbb{S}_b^+,\ b \in \mathbb{B}\right\}$ obtained via Monte Carlo simulations in STAGE I and the selected CAC threshold $a_{\max}$, it is the objective of STAGE II to derive the desired performance measures by incorporating these interference-related aspects with the traffic dynamics in a Markovian analysis. Interpreting the outage probability as the BLER, the expected effective DSCH throughput (in kbits/s) per data call offered by NodeB $b \in \mathbb{B}$ in system state $\mathbf{a} \in \mathbb{S}_b^+$ is given by

$$\beta_b(\mathbf{a}) \equiv a_b^{-1}\left(1 - P_b(\mathbf{a})\right) R_{\text{DSCH}}.$$

Along with the Poisson call arrival rates and the exponentially distributed call sizes, the system evolution can be described by the irreducible continuous-time Markov chain $(\mathbf{A}(t))_{t\geq 0} \equiv (A_1(t), A_2(t), \cdots, A_B(t))_{t\geq 0}$ with states denoted $\mathbf{a} \in \mathbb{S}$. The Markov chain's infinitesimal generator, denoted $\mathcal{Q}$, is defined by

$$\mathcal{Q}(\mathbf{a}, \mathbf{a}') = \begin{cases} \lambda_b\left(1 - \mathbf{P}_f^{\mathcal{G}}\right) & \text{if } \mathbf{a}' = \mathbf{a} + \mathbf{e}_b \\ \\ \mu a_b \beta_b(\mathbf{a}) & \text{if } \mathbf{a}' = \mathbf{a} - \mathbf{e}_b \end{cases}$$

for $\mathbf{a}, \mathbf{a}' \in \mathbb{S}$, where $\mathbf{e}_b$ is the B-dimensional vector with all zeros, except a one on position b, and $\mathbf{P}_f^{\mathcal{G}}$ denotes the fraction of calls that fails to satisfy the minimum path gain requirement (6.2). All other non-diagonal entries of $\mathcal{Q}$ are 0, while the diagonal entries are such that all rows of $\mathcal{Q}$ sum up to 0. Since the finite state space Markov chain $(\mathbf{A}(t))_{t\geq 0}$ is irreducible, a unique probability vector $\boldsymbol{\pi}$ exists that satisfies the system of global balance equations $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0}$ with $\mathbf{0}$ the vector with all entries zero. For $B > 1$, no closed-form expression for $\boldsymbol{\pi}$ is known, due to the influence of $a_{b'}$, $b' \neq b$, on the DSCH throughput at NodeB b. Therefore, numerical methods, e.g. the successive overrelaxation method (see [213]) are applied to determine the equilibrium distribution.

PERFORMANCE MEASURES

A number of basic performance measures can be obtained directly from the equilibrium distribution. From a system's perspective, the (cell-specific) expected channel utilisation expresses the achieved resource efficiency, while the Grade Of Service is readily determined in terms of the overall (cell-specific) fresh call blocking probabilities (using the PASTA property [224]):

$$\mathbf{P}_{f,b} \equiv \mathbf{P}_f^{\mathcal{G}} + \left(1 - \mathbf{P}_f^{\mathcal{G}}\right) \sum_{\{\mathbf{a} \in \mathbb{S}:\; a_b = a_{\max}\}} \pi\left(\mathbf{a}\right),$$

and

$$\mathbf{P}_f \equiv \sum_{b \in \mathbb{B}} \left(\frac{\lambda_b}{\sum\limits_{b' \in \mathbb{B}} \lambda_{b'}} \right) \mathbf{P}_{f,b},$$

which include the effects of *both* CAC conditions. Of principal relevance in the presented investigation is the experienced data QOS, expressed as the (cell-specific) expected transfer time $\mathbf{T}$ ($\mathbf{T}_b$, $b \in \mathbb{B}$) of a data call, which is readily derived applying Little's formula:

$$\mathbf{T}_b \equiv \frac{\mathbf{N}_b}{\lambda_b \left(1 - \mathbf{P}_{f,b}\right)} = \frac{\sum\limits_{\mathbf{a} \in \mathbb{S}} a_b\, \pi\left(\mathbf{a}\right)}{\lambda_b \left(1 - \mathbf{P}_{f,b}\right)}, \quad b \in \mathbb{B},$$

and

$$\mathbf{T} \equiv \sum_{b \in \mathbb{B}} \left(\frac{\lambda_b \left(1 - \mathbf{P}_{f,b}\right)}{\sum\limits_{b' \in \mathbb{B}} \lambda_{b'} \left(1 - \mathbf{P}_{f,b'}\right)} \right) \mathbf{T}_b = \frac{\sum\limits_{\mathbf{a} \in \mathbb{S}} \left(\sum\limits_{b \in \mathbb{B}} a_b \right) \pi\left(\mathbf{a}\right)}{\left(\sum\limits_{b \in \mathbb{B}} \lambda_b \right) \left(1 - \mathbf{P}_f\right)},$$

where $\mathbf{N}_b$ denotes the expected number of data calls in cell $b \in \mathbb{B}$.

Aside from the above performance measures which can be obtained directly from the Markov chain's equilibrium distribution, the expected transfer time $\widehat{\tau}_{b,\mathbf{a}}(x)$ of a data call admitted to NodeB $b \in \mathbb{B}$ can be determined conditional on the state $\mathbf{a} \in \mathbb{S}_b^+$ of the system upon call arrival (where a_b includes the new data call) and on the data call size $x \in \mathbb{R}^+$ (in kbits). Since the derivation follows the same analytical lines as the conditional analysis in Chapters 2 and 3, we merely state the result here. Let $\mathcal{Q}_b^\star$ denote the infinitesimal generator of the modified version of the original Markov chain $(\mathbf{A}(t))_{t\geq 0}$ characterised by the presence of one permanent data call in cell $b \in \mathbb{B}$, i.e.

$$\mathcal{Q}_b^\star(\mathbf{a},\mathbf{a}') = \begin{cases} \lambda_{b'}\left(1 - \mathbf{P}_f^{\mathcal{G}}\right) & \text{if } \mathbf{a}' = \mathbf{a} + \mathbf{e}_{b'}, \\ \\ \mu\left(a_b - 1\right)\beta_b\left(\mathbf{a}\right) & \text{if } \mathbf{a}' = \mathbf{a} - \mathbf{e}_b, \\ \\ \mu a_{b'}\beta_{b'}\left(\mathbf{a}\right) & \text{if } \mathbf{a}' = \mathbf{a} - \mathbf{e}_{b'} \text{ with } b' \neq b, \end{cases}$$

for $\mathbf{a}, \mathbf{a}' \in \mathbb{S}_b^+$. All other non-diagonal entries of $\mathcal{Q}_b^\star$ are 0, while the diagonal entries are such that all rows of $\mathcal{Q}_b^\star$ sum up to 0. The permanent data call, i.e. the tagged call whose transfer time is to be determined, never leaves the system, but shares in the varying DSCH throughput as if it were finite. Let $\boldsymbol{\pi}_b^\star$ be the stationary probability distribution vector corresponding to the modified Markov chain, i.e. $\boldsymbol{\pi}_b^\star \mathcal{Q}_b^\star = \mathbf{0}$. Furthermore, let $\mathcal{B}_b \equiv \mathit{diag}(\beta_b(\mathbf{a}),\ \mathbf{a} \in \mathbb{S}_b^+)$ denote the diagonal matrix of expected effective DSCH throughputs per data call. The closed-form solution for $\widehat{\boldsymbol{\tau}}_b(x) \equiv \left(\widehat{\tau}_{b,\mathbf{a}}(x),\ \mathbf{a} \in \mathbb{S}_b^+\right)$ is given by

$$\widehat{\boldsymbol{\tau}}_b(x) = \frac{x}{\boldsymbol{\pi}_b^\star \mathcal{B}_b \mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{x\mathcal{B}_b^{-1}\mathcal{Q}_b^\star\right\}\right]\boldsymbol{\gamma}_b,$$

where $\boldsymbol{\gamma}_b \equiv \left(\gamma_b\left(\mathbf{a}\right),\ \mathbf{a} \in \mathbb{S}_b^+\right)$ is the unique solution to the system of linear equations

$$\begin{aligned} \mathcal{Q}_b^\star \boldsymbol{\gamma}_b &= \frac{\mathcal{B}_b \mathbf{1}}{\boldsymbol{\pi}_b^\star \mathcal{B}_b \mathbf{1}} - \mathbf{1}, \\ \boldsymbol{\pi}_b^\star \mathcal{B}_b \boldsymbol{\gamma}_b &= 0. \end{aligned}$$

The conditional expected transfer time of a data call of size x served by NodeB b is then given by

$$\mathbf{T}_b(x) \equiv \sum_{\mathbf{a}\in\mathbb{S}_b^+} \widehat{\tau}_{b,\mathbf{a}}(x) \left(\frac{\pi(\mathbf{a}-\mathbf{e}_b)}{\sum\limits_{\mathbf{a}'\in\mathbb{S}_b^+} \pi(\mathbf{a}'-\mathbf{e}_b)} \right),$$

where $\mathbf{e}_b$ is the B-dimensional vector with a one on the b^{th} position and zeros elsewhere.

OTHER EXPERIMENTS

The performance analyses for EXPERIMENTS 0 and 1A are similar to those of EXPERIMENTS 2 and 3 presented above. The only difference is that the expected effective DSCH throughput per data call $\beta_1(a_1)$ offered by NodeB 1 ($\mathbb{B}=\{1\}$) in system state $a_1 \in \mathbb{S}_1^+$, is now given by

$$\beta_1(a_1) \equiv \begin{cases} a_1^{-1} R_{\text{DSCH}} & (\text{EXPERIMENT } 0), \\ \\ a_1^{-1}\left(1-P_1^{\bullet}(1)\right) R_{\text{DSCH}} & (\text{EXPERIMENT } 1\text{A}). \end{cases}$$

While the multi-NodeB EXPERIMENTS 2 and 3 require numerical procedures to obtain the Markov chain's equilibrium distribution and thus the desired performance measures, equivalent measures for the single-NodeB EXPERIMENTS 0, 1A and 1B can be obtained in closed-form. The model of EXPERIMENT 1B is an $M/G/1/a_{\max}/GPS$ queueing model with state-dependent aggregate service rates given by $a_1\beta(a_1) \equiv (1-P_1(a_1)) R_{\text{DSCH}}$, $a_1 = 1,\cdots,a_{\max}$. Of primary interest is the (conditional) expected transfer time $\mathbf{T}$ ($\mathbf{T}(x)$), which for the $M/G/1/a_{\max}/GPS$ model is given by

$$\mathbf{T}(x) \equiv \frac{x}{R_{\text{DSCH}}} \sum_{a_1=0}^{a_{\max}} \frac{a_1 \pi(a_1)}{\rho\left(1-\mathbf{P}_f^{\mathcal{G}}\right)\left(1-\pi(a_{\max})\right)},$$

and

$$\mathbf{T} \equiv \sum_{a_1=0}^{a_{\max}} \frac{a_1 \pi(a_1)}{\lambda \left(1 - \mathbf{P}_f^{\mathcal{G}}\right) \left(1 - \pi(a_{\max})\right)},$$

where

$$\pi(a_1) = \frac{\left(\rho \left(1 - \mathbf{P}_f^{\mathcal{G}}\right)\right)^{a_1} \phi(a_1)}{\sum_{a_1'=0}^{a_{\max}} \left(\rho \left(1 - \mathbf{P}_f^{\mathcal{G}}\right)\right)^{a_1'} \phi(a_1')},$$

$a_1 = 0, \cdots, a_{\max}$, is the model's equilibrium distribution, with

$$\phi(a_1) \equiv \left(\prod_{a_1'=1}^{a_1} \frac{a_1' \beta_1(a_1')}{R_{\text{DSCH}}} \right)^{-1} = \left(\prod_{a_1'=1}^{a_1} \left(1 - P_1(a_1')\right) \right)^{-1},$$

$a_1 = 0, \cdots, a_{\max}$, where $\phi(0) \equiv 1$ by convention. These expressions have been shown to be insensitive to the data call size distribution except for its mean [54]. The models of EXPERIMENTS 0 and 1A are instances of the basic $M/G/1/PS/a_{\max}$ queueing model with fixed aggregate service rates $a_1\beta_1(a_1)$ equal to R_{DSCH} and $(1 - P_1^{\bullet}(1)) R_{\text{DSCH}}$, respectively, and are thus special cases of the above-mentioned queueing model used to conduct EXPERIMENT 1B.

6.7. NUMERICAL RESULTS

Following the evaluation approach outlined in Section 6.5, a number of insightful numerical experiments are now presented to assess the 'Round Robin effect' and the different 'interference effects' caused by the DSCH's self-interference, the presence of A-DCHs and the influence from adjacent NodeBs. Furthermore, the impact of the downlink orthogonality factor ω and the (uniform) traffic load $\rho_b = \rho$, $b \in \mathbb{B}$, is determined. The assumed parameter settings are summarised in Table 6.1 below. The values given for the maximum transmission power $p_{\max}$ and thermal noise level ν correspond to 42 dBm and −99.157 dBm, respectively, where the effective thermal noise level ν equals $R_c \cdot k \cdot T \cdot N_f$, with $R_c = 3840$ kchips/s, $k \approx 1.38\,10^{-23}$ J/K is Boltzmann's constant, $T = 290$ K is the considered absolute temperature and $N_f = 9$

dB is the assumed receiver noise figure (see e.g. [141]). The assumed hexagonal cell radius of $1/\sqrt{3} \approx 0.577$ km corresponds to an inter-NodeB distance of precisely 1 km. The downlink orthogonality factors are taken from [78] ($\omega = 0.06$, indoor office test environment) and [167] ($\omega = 0.65$, typical urban channel) in order to consider two very distinct yet realistic alternatives.

Table 6.1 Settings of the system, propagation and traffic model parameters for the numerical evaluations.

SYSTEM MODEL			TRAFFIC MODEL		
B	$\in \{1,2,3\}$	NodeBs	μ^{-1}	320	kbits
radius	$1/\sqrt{3}$	km	λ	$\in [0, 2\mu R_{\text{DSCH}}]$	call/s
R_c	3840	kchips/s	$a_{\max}$	17	calls
R_{DSCH}	1024	kbits/s	PROPAGATION MODEL		
γ_{DSCH}	4	dB	η_{basic}	137.744	dB
$R_{\text{A-DCH}}$	3.4	kbits/s	ς	3.523	-
$\gamma_{\text{A-DCH}}$	7	dB	a, b	$1/\sqrt{2}$	-
$p_{\max}$	15.849	Watt	σ_{S}	8	dB
ν	$1.214\ 10^{-13}$	Watt	ω	$\in \{0.06, 0.65\}$	-

As with the performance analysis, the numerical experiments are presented in two separate stages: we first investigate the outage probability caused by the different interference aspects (STAGE I), followed by the transfer time analysis that includes the traffic dynamics in STAGE II.

6.7.1. INTERFERENCE ASPECTS

The objective of STAGE I is to determine the set of outage probability functions $\{P_b(\mathbf{a}), \mathbf{a} \in \mathbb{S}_b^+, b \in \mathbb{B}\}$, conditional on the path gain-based CAC criterion, by means of Monte Carlo simulations. The number of snapshots was taken to be $K = 100000$ to ensure that the relative precision of the constructed 95% confidence intervals is no worse than 5%. Consider first EXPERIMENTs 2 and 3. Given the network symmetry implied by the choice of $B \in \{2, 3\}$ (see e.g. Figure 6.2), it holds that

$$\begin{cases} P_1(a_1, a_2) = P_2(a_2, a_1) & \text{for } B = 2, \\ \\ P_1(a_1, a_2, a_3) = P_2(a_2, a_1, a_3) = P_3(a_3, a_1, a_2) & \text{for } B = 3, \end{cases}$$

and hence it suffices to determine only $\left\{P_1(\mathbf{a}),\ \mathbf{a} \in \mathbb{S}_1^+\right\}$.

For $B = 2$, $R_{\text{DSCH}} = 1024$ and $\omega \in \{0.06, 0.65\}$, Figure 6.4 shows the outage probability $P_1(\mathbf{a})$ as a function of $\mathbf{a} \in \mathbb{S}_1^+$. The accompanying Figure 6.5 shows the same numerical results but now in the form of iso-outage curves, a representation that may be useful to determine an appropriate CAC region. Aside from the trivial observation that $P_1(\mathbf{a})$ is increasing in both a_1 and a_2, the figure allows some additional insightful observations. There is a significant increase in $P_1(\mathbf{a})$ from $a_2 = 0$ to $a_2 = 1$, due to the activation of the DSCH in the adjacent cell, whose high data rate and the correspondingly low processing gain, requires a high transmission power and thus induces a large increase in the interference level experienced in the reference cell. In Figure 6.5 this sudden increase is apparent from the iso-outage curves' near alignment with the horizontal axes. As a_2 becomes larger, $P_1(\mathbf{a})$ increases with a much less dramatic (slightly positive) slope, since the raised interference is now only caused by the additional low data rate A-DCHs, requiring relatively low transmission power levels. Comparing both charts, note that for $\omega = 0.65$, $P_1(\mathbf{a})$ shows a much greater dependency on a_1, due to the smaller orthogonality gain in the reference cell. As a consequence, the relatively high transmission powers thus required for greater a_1 and the resulting interference levels induce higher transmission powers at NodeB 2, establishing an increasing dependency between both NodeBs. Although this effect is present in both charts, it is particularly visible in the chart for $\omega = 0.65$.

With regard to the absolute values of the outage probabilities, note that for $\omega = 0.06$ (0.65), $P_1(\mathbf{a})$ varies between 0.0360 (0.0510) for $\mathbf{a} = (1, 0)$ and 0.0744 (0.1477) for $\mathbf{a} = (30, 30)$. As a visual presentation of the outage probabilities is not possible for $B = 3$, it is noted for comparison that in a network of three cells $P_1(\mathbf{a})$ varies between 0.0343 (0.0497) for $\mathbf{a} = (1, 0, 0)$ and 0.1015 (0.2230) for $\mathbf{a} = (30, 30, 30)$, for downlink orthogonality factor $\omega = 0.06$ (0.65). Observe that $P_1(1, 0, 0)$ $(B = 3)$ is slightly lower than $P_1(1, 0)$ $(B = 2)$ since the additional NodeB may relieve NodeB 1 of unfavourably (e.g. due to shadowing) located terminals (enhanced diversity gain),

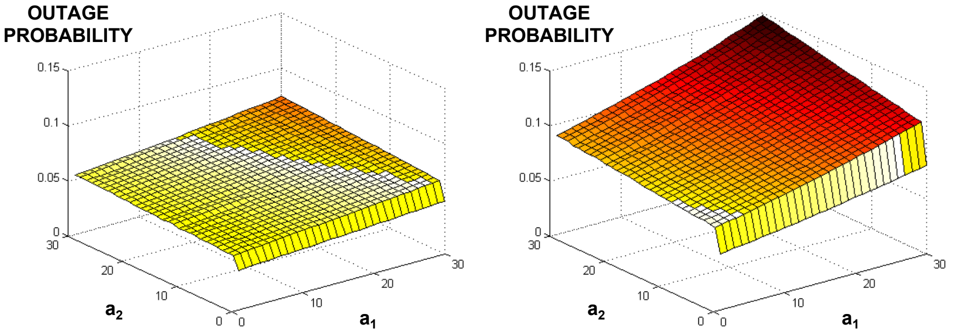


Figure 6.4 EXPERIMENT 2: outage probabilities $P_1(a_1, a_2)$ as a function of the number of data calls present in cell 1 (a_1) and 2 (a_2) for the case of $\omega = 0.06$ (left) and $\omega = 0.65$ (right).

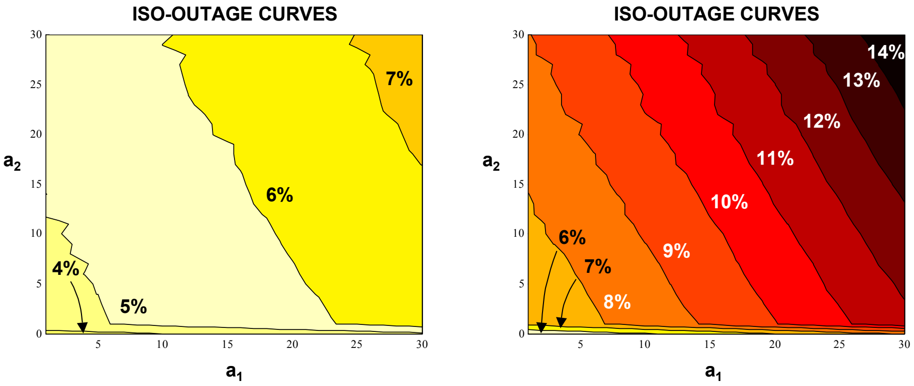
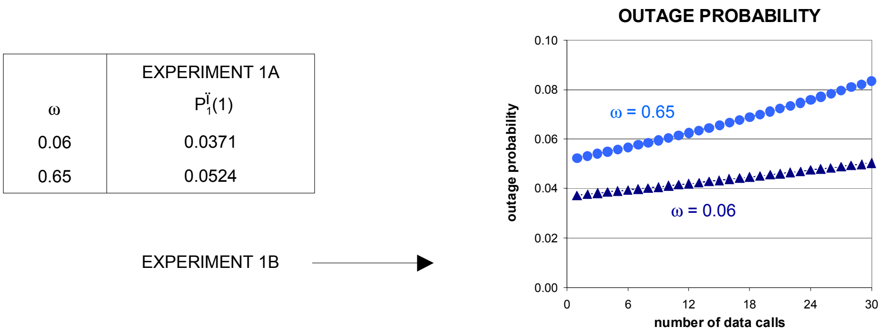


Figure 6.5 EXPERIMENT 2: iso-outage curves in terms of the number of data calls present in cell 1 (a_1) and 2 (a_2) for the case of $\omega = 0.06$ (left) and $\omega = 0.65$ (right).

rather than allowing a poor QOS for the unfortunate call. Trivially, $P_1(30, 30, 30)$ $(B = 3)$ exceeds $P_1(30, 30)$ $(B = 2)$ significantly, due to the additional interference caused by NodeB 3's DSCH and A-DCHs.

The outage probabilities for EXPERIMENTS 1A and 1B are given in Figure 6.6. Observe that in EXPERIMENT 1B, $P_1(1)$ is equal to 0.0373 (0.0524) for $\omega = 0.06$ (0.65), which implies that even in a single cell case serving a single randomly located data call, a DSCH at a rate of 1 Mbits/s may suffer a significant expected outage probability (or BLER), due to the interference generated by the single A-DCH and the lack of perfect signal orthogonality due to multipath fading. In comparison with the corresponding values obtained for the experiments *with* A-DCHs, the outage probability $P_1^{\bullet}(1)$ for EXPERIMENT 1A (*without* A-DCHs) is slightly lower due to the reduced interference levels. The above-mentioned diversity gain observed in larger networks, induces that e.g. $P_1(a_1, 0, 0) < P_1(a_1, 0) < P_1(a_1)$, $a_1 = 1, 2, \cdots, a_{\max}$ (EXPERIMENTS 3, 2, 1B).



	EXPERIMENT 1A
ω	$P^{\text{Ï}}_1(1)$
0.06	0.0371
0.65	0.0524

Figure 6.6 EXPERIMENTS 1A and 1B: outage probabilities $P_1^{\bullet}(1)$ and $P_1(a_1)$ as a function of the number of data calls a_1 present in the considered cell, respectively, for the cases of $\omega \in \{0.06, 0.65\}$.

As the STAGE I Monte Carlo simulations are used to derive an outage probability function, *conditional* on the compliance with the path gain-based CAC condition, the fraction of data calls $\mathbf{P}_f^{\mathcal{G}}$ that is rejected by this CAC scheme is also obtained. Table 6.2 contains $\mathbf{P}_f^{\mathcal{G}}$ for all considered experiments. Observe that $\mathbf{P}_f^{\mathcal{G}}$ increases in ω due to a reduced orthogonality gain, decreases in the number of NodeBs due to an improved diversity gain, and increases slightly if the A-DCHs are included due to both the associated interference and the reduced NodeB power budget that can be assigned to the actual DSCH transfer.

As stated above, the outage probability curves can be useful to derive the secondary CAC threshold $a_{\max}$ and thus pose an upper bound on the outage probability. For the STAGE II experiments, we have assumed a typical outage requirement of less than 8%, which in a network of 3 NodeBs and $\omega = 0.06$ implies a CAC threshold of $a_{\max} = 17$ (in a network of 1 (2) NodeB(s) with the same multipath conditions this CAC threshold limits the outage probability to 4.42% (6.12%)). Whereas in practice,

Table 6.2 Fraction $\mathbf{P}_f^{\mathcal{G}}$ of data calls that is rejected due to the path gain-based CAC criterion.

ω	EXPERIMENT 0	EXPERIMENT 1A	EXPERIMENT 1B	EXPERIMENT 2	EXPERIMENT 3
0.06	0.0000%	0.9380%	0.9475%	0.6814%	0.5342%
0.65	0.0000%	1.8658%	1.8927%	1.4450%	1.1784%

other CAC thresholds (possibly taking other NodeBs' actual loading into account) may of course be set, depending on the specific propagation environment, the network size and the operator's policy, our qualitative comparisons require a uniform CAC threshold in all considered experiments.

6.7.2. TRAFFIC DYNAMICS

For each of the experiments, the required BLER functions determined via Monte Carlo simulations in STAGE I are used to determine the state-dependent expected throughput in the STAGE II Markov chain model that captures the traffic dynamics. The performance analyses have been presented in Section 6.6 above.

For $R_{\text{DSCH}} = 1024$, $\omega \in \{0.06, 0.65\}$ and $\rho \in (0, 2]$, Figure 6.7 shows the expected transfer times (in seconds) versus the data traffic load for all five experiments. At this point it is stressed that the traffic load is implicitly defined in terms of a uniform data call arrival rate λ for a *given* call size average μ^{-1}. While for EXPERIMENTS 0 and 1A the presented results are sensitive to λ and μ only through their ratio ρ, this is certainly not the case for EXPERIMENTS 1B, 2 and 3, as in these experiments the *number* of present data calls affects the *aggregate* service rate due to the interference generated to maintain the A-DCHs.

The general form of the curves shows an exponential QOS degradation for $\rho \in (0, 1]$, primarily due to the increase in the number of data calls sharing the DSCH resources, which is followed by a QOS stabilisation effectuated by the Call Admission Control scheme ($a_{\max}$). It is the relative performance of the different experiments and downlink orthogonality factors that is of principal interest. In order to assess the different 'interference effects', the lower curve (EXPERIMENT 0) serves as a reference for the other curves as it represents the most basic scenario (capturing the 'Round Robin effect' only). For both values of ω, observe that in the single cell case the overall interference effect is captured for the better part by EXPERIMENT 1A, indicating that

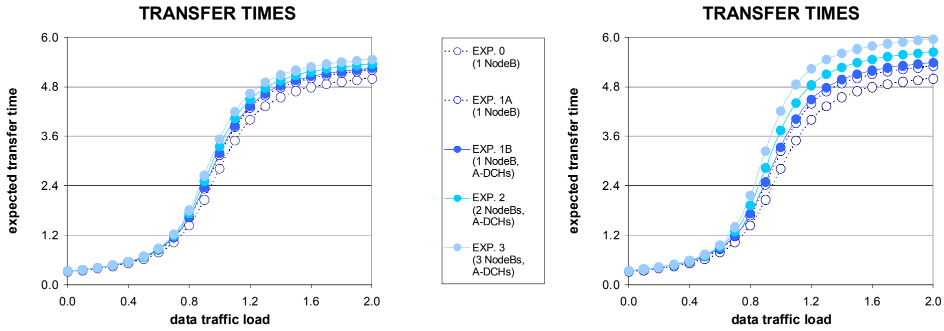


Figure 6.7 Expected transfer time of a data call versus the data traffic load ρ for the cases of $\omega = 0.06$ (left) and $\omega = 0.65$ (right) and all five considered experiments.

the inclusion of path gain-based CAC, the impact of the power budget limitation and primarily the DSCH's self-interference to the reference case, is substantially more significant than the further inclusion of the A-DCHs in EXPERIMENT 1B.

An extension of the system with additional NodeBs in EXPERIMENTS 2 and 3 induces a further significant QOS degradation. Apparently (and expectedly), the raised interference levels due to the extra DSCH(s) and A-DCHs outweigh the enhanced diversity gain. Concentrating on the influence of ω, observe that not only the *absolute* QOS is worsened by a smaller degree of orthogonality (higher ω), but also the *relative* QOS degradation due to the addition of extra NodeBs, is largest for the case of $\omega = 0.65$, even though a low ω indicates a *relatively* large intercellular interference contribution. Apparently, the fact that, in a myopic sense, a higher ω leads to higher transmission powers within each cell separately, outweighs the relatively small intercellular interference coupling. Although this is not demonstrated, it is important to note that the Call Admission Control threshold $a_{\max}$ may also strongly influence the relative interference impact of the DSCHs and A-DCHs in adjacent cells. For instance, a more stringent Call Admission Control (lower $a_{\max}$) reduces both the impact of the A-DCHs and the activity level of the DSCHs.

Figure 6.8 presents the conditional expected transfer time of a data call of size $x \in (0, 500]$ kbits for $\rho = 1$. While the graphs for EXPERIMENTS 0, 1A and 1B are straight lines, those for EXPERIMENTS 2 and 3 are slightly concave (see also Chapter 4 and [174]) due to the random variations in the effective aggregate DSCH throughput caused by the dynamics in adjacent cells. Observe that the ordering of the curves corresponding to the different experiments agrees with those in Figure 6.7 (for $\rho = 1$), while the impact of ω on the relative performance of the different experiments is also similar. As the different curves are (approximately) straight lines, the 'relative transfer time mark-up' of the different model extensions incorporated in the experiments is (approximately) constant in the data call size x. For instance, for $\omega = 0.65$, the conditional expected transfer times in EXPERIMENTS 1A, 1B, 2 and 3, are approximately 15.30%, 18.67%, 37.28% and 49.30% higher, respectively, than those in EXPERIMENT 0.

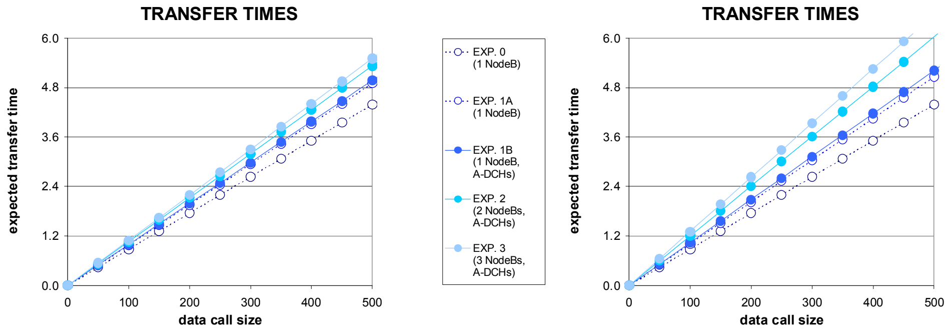


Figure 6.8 Conditional expected transfer time of a data call versus its size for the cases of $\omega = 0.06$ (left) and $\omega = 0.65$ (right) and all five considered experiments.

The numerical section concludes with a comparison of the results obtained using the semi-analytical two-stage approach and those obtained by means of direct time-consuming dynamic simulations of the considered system. Figure 6.9 (left) shows the expected transfer time curves for the case of $\omega = 0.65$. Note that the curve associated with EXPERIMENT 0 is identical to the corresponding curves in Figure 6.7. Observe first that the absolute QOS levels obtained in the simulation experiments are worse than the corresponding analytically obtained values. Three distinct causes can be identified for this discrepancy. *(i)* The simulations are run with a fixed 10 ms (UMTS time frame) heartbeat which typically requires padding of each call's final data transport block and thus corresponds with a waste of capacity. *(ii)* The second stage in the analytical approach basically assumes all terminals to have an 'average location', whereas the dynamic simulations actually take the location variability into account by randomly sampling the location corresponding to each generated data call. *(iii)* The analytical approach implicitly assumes all competing data calls to be homogeneously distributed over the corresponding cells.

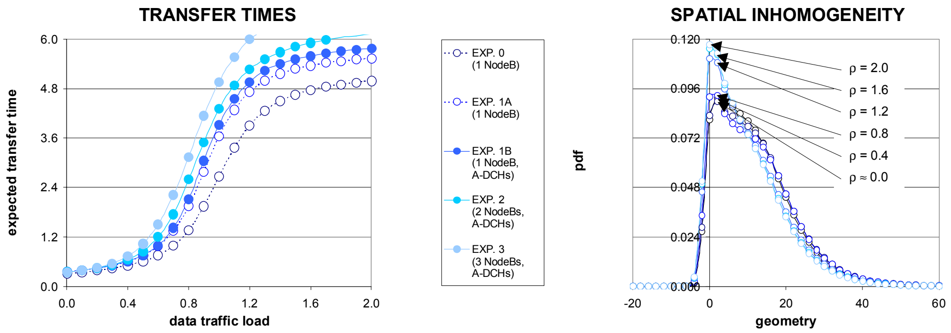


Figure 6.9 Dynamic simulation results. The left chart shows the expected transfer time of a data call versus the data traffic load ρ for the case of $\omega = 0.65$ and all five considered experiments. The right chart indicates the induced spatial inhomogeneity by the PDFs of the UE's geometry for $\omega = 0.65$ and $\rho \in \{\approx 0.0, 0.4, 0.8, 1.2, 1.6, 2.0\}$.

However, the simulation results in Figure 6.9 (right) show that even though the arrival process is spatially homogeneous, the spatial distribution of the calls present at any given time is typically skewed, due to the phenomenon that a data call nearer to its serving NodeB experiences fewer block errors and thus sooner completes its intended transfer and departs from the system. The chart demonstrates this using the PDFs of the so-called call *geometry*, which for call m is defined as

$$\Omega_m \equiv \frac{p_{\max}\mathcal{G}_{mb_m}}{\sum\limits_{b\neq b_m} p_{\max}\mathcal{G}_{mb}+\nu},$$

and captures the relative location and includes the effects of shadowing. A large value of Ω_m indicates that call m resides relatively close to its serving NodeB, while a small value of Ω_m indicates that call m is typically located in the intersection of the service areas of neighbouring NodeBs. The geometry is generally expressed in dB. The depicted PDF associates with each geometry value the relative aggregate transfer time that is experienced by data calls with this geometry value. As the chart shows, a heavier traffic load 'glues' terminals to the cells' edges, corresponding with a more severe spatial traffic inhomogeneity. Compared to the analytical results, this spatial skewness causes an additional interference increase, and thus a further QOS degradation. We note that a slight degree of spatial skewness occurs even for a negligible traffic load, since a NodeB's power budget may be insufficient to overpower the thermal noise level and self-interference at a remote terminal receiver in the presence of highly variable Rayleigh fading.

Aside from these absolute discrepancies, observe that the *qualitative* trends of the semi-analytical approach appear to be in line with those indicated by the simulation results (see Figure 6.7 (right)).

6.8. CONCLUDING REMARKS

We have presented a semi-analytical performance analysis of data transfer over Downlink Shared CHannels in UMTS networks. Following an insightful two-stage approach to segregate the interference aspects from the traffic dynamics, we have decomposed the general performance model into simpler experiments, in order to provide qualitative insight in the influence of different system aspects on the data QOS. In particular, the analysis has demonstrated that while a heavier data traffic load not only implies a greater competition for DSCH resources and thus longer transfer times ('Round Robin effect'), it also implies a higher interference level due to the greater number of A-DCHs that must be maintained for signalling purposes, which causes a higher BLock Error Rate and thus a lower effective aggregate DSCH throughput ('interference effect'). In summary: *the greater the demand for service, the smaller the aggregate service capacity*. The principal advantages of the presented two-stage performance evaluation method are its relative swiftness, the benefits gained from the method's first stage, e.g. the possibility to derive CAC rules, and the delivered insight into the conditional data QOS given a call's size.

Among the observations from the presented numerical experiments, we note that the impact of the DSCHs' self-orthogonality on the data QOS dominates the impact of the interference from the A-DCHs, thus validating the assumption often made in related investigations (e.g. Chapter 7 and [13, 137]) to disregard the A-DCHs. The results further illustrate that the raised interference caused by the inclusion of a(n additional) adjacent NodeB leads to a QOS degradation that is slightly smaller than that caused by the inclusion of intra-cellular interference in a basic UMTS model. Although this is not numerically supported, the impact of including non-adjacent (e.g. second tier) NodeBs is typically much smaller. The influence of the downlink orthogonality factor was demonstrated to be fairly significant, with a lower degree of orthogonality leading to worse data QOS.

A comparison with direct dynamic simulations revealed that although absolute QOS levels differ, the principal *qualitative* conclusions of the presented investigations are pleasingly backed. An important reason for the discrepancy in the absolute QOS performance between the presented semi-analytical approach and direct simulations, has been argued to be related to the spatial terminal distribution, whose skewness was argued and demonstrated to increase with a heavier data traffic load. Hence potential improvements of the analytical evaluation approach should attempt to capture this spatial skewness phenomenon. One approach could be to specify terminal locations with a granularity finer than per cell, e.g. by partitioning each cell in (concentric) zones, which would effectively translate into a higher-dimensional state space and thus potential computational problems in the second stage of the analysis. Other potentially rewarding approaches include an intelligent adjustment of the spatially uniform terminal distribution in the STAGE I Monte Carlo analysis, or the application of a scheduling discipline which establishes the sort of fairness that largely preserves the spatially uniform terminal distribution. Whether a scheduling scheme with such a homogeneity preserving objective is also adequate in terms of resource efficiency and the delivered QOS, is assessed with a broader scope in the next chapter for an integrated services UMTS/HSDPA network.

CHAPTER 7

FAIR ADAPTIVE SCHEDULING IN AN INTEGRATED SERVICES UMTS NETWORK

IN view of the inherent scarcity of radio capacity, the efficient sharing of resources in an integrated services UMTS network is as principal a challenge as it is for e.g. GSM/GPRS networks. The involved control operations are however fundamentally different due to the differences in radio access technology. Whereas GSM's FD/TDMA-based radio interface as considered in Chapters 2 and 3 allows an unambiguous definition of capacity, i.e. in number of available traffic channels, the available capacity in a CDMA-based network is rather vaguely determined by a.o. the interference levels and the location of the communicating entities. If a UMTS radio link consumes more radio resources than available, however, e.g. as a consequence of terminal mobility, this results in a *graceful* degradation of all affected radio links, while for a similar scenario in a GSM network the required capacity to maintain the considered radio link would be simply unavailable, resulting in a dropped call.

A common characteristic of the service integration policies investigated for GSM/HSCSD (Chapter 2) and GSM/GPRS networks (Chapter 3), respectively, was the *fair* sharing of available resources by all data flows (within the same priority class). Extending this fairness objective to the UMTS domain is non-trivial, as it now makes an important difference whether fairness is measured by the assigned resources or by the experienced data throughput. As a consequence of the CDMA access technology, a remote terminal typically experiences a lower downlink data throughput than a terminal close to a NodeB, even if the same amount of transmission power is assigned to the radio link. This phenomenon is primarily due to the applied universal frequency reuse and the consequentially significant impact of the experienced interference levels, and hence a terminal's location, on the QOS, but is further amplified by the link adaptation mechanisms developed for the HSDPA system upgrade. We note that as e.g. EDGE-enhanced networks also feature a link adaptation mechanism with a fairly wide dynamic range of operation, the above-mentioned distinction in fairness objectives also applies to advanced FD/TDMA-based 2G systems.

The research presented in this chapter concentrates on the efficient integration of prioritised speech and delay-tolerant data traffic in a UMTS network. *Adaptive scheduling* is the control mechanism applied to achieve both resource efficiency and fairness, which is a combined responsibility of the NodeB and RNC, and is defined to comprise of two principal tasks (see also Chapter 1):

RATE CONTROL: The rate control mechanism dynamically up/downgrades the data transport channels in accordance with variations in prioritised speech traffic loads. Although perhaps slightly confusing given the mechanism's name, in the context of the proposed HSDPA improvements of the UMTS standard, our investigations also include a parameterisation of the downlink shared data transport channels by an assigned transmission *power*, rather than a transfer *rate*.

PACKET SCHEDULING: Packet scheduling refers to the controlled time-multiplexing of the data flows in order to optimise resource efficiency while satisfying the calls' QOS requirements and providing some sort of fairness.

These mechanisms differ both in their scope and the operational time scale. Rate control typically has a broader spatial scope and a larger operational time scale (seconds). Although the UMTS specifications allow e.g. the DSCH to be up- and downgraded on a time frame (10 ms) basis [1], in practice it is unlikely that rate control will operate on such a small time scale for reasons of stability. In contrast, packet scheduling is typically of a rather myopic nature and operates on the smaller millisecond time scale.

Whereas on the uplink common channels (RACH, CPCH) a contention resolution scheme such as slotted ALOHA, Carrier Sense Multiple Access, or Packet Reservation Multiple Access, is typically deployed to coordinate transmissions, the packet scheduling of data bursts on downlink common channels is under direct control of the NodeB or RNC, which is typically proposed to adopt e.g. a Round Robin, Weighted Fair Queueing, Earliest Deadline First, or waterfilling scheme (see e.g. [107]). While in a basic UMTS network the data rate or transmission power that is available to the packet scheduler on these common transport channels are manually set and adjusted by the network operator, an enhanced UMTS network may deploy rate control schemes for both up- and downlink common channels, in order to e.g. improve data throughputs under a varying load of prioritised speech traffic or to dynamically shift resources in support of data traffic hot spots. As we will demonstrate, by appropriately down- or upgrading the data transport channels in incidences of heavy or light speech traffic, respectively, the performance of both service types can be improved.

With regard to the dynamic shifting of resources towards traffic hot spots as a form of rate control, observe that in e.g. speech-only CDMA-based radio networks, capacity moves towards busier areas in a natural manner, since a local increase of the carried load reduces the capacity in the immediate surroundings. This CDMA-inherent phenomenon, which bears some equivalence to dynamic channel allocation in FDMA-based networks (e.g. GSM), holds only to a limited extent when data transfers are scheduled over a fixed-rate common channel (e.g. DSCH). Since the aggregate bit rate of a common channel is typically limited, a local increase in the carried data traffic load translates only partially into an increased interference level and thus a reduced capacity in the surrounding area. In particular if the local data traffic load is relatively heavy with respect to the common channel's aggregate bit rate, an increase in the data traffic load primarily affects the experienced QOS, rather than the transferred bit rate and thus the generated interference. Adaptive rate control may then be effective to spatially shift capacity in order to enhance data throughputs and achieve fairness.

The outline of the chapter is as follows. Section 7.1 reviews the relevant literature, followed by a statement of contribution in Section 7.2. As the presented work partially relies on the deployment of HSDPA, an overview of this technological upgrade of UMTS is given in Section 7.3. Section 7.4 then sets the framework for the performance evaluation of the different scheduling schemes which are specified in Section 7.5. The performance evaluation methods are outlined in Section 7.6, while the subsequent Section 7.7 presents a numerical study and discusses the results. Section 7.8 ends this chapter with some concluding remarks.

7.1. LITERATURE

The presented review concentrates only on those rare references that investigate the rather advanced mechanism of adaptive rate control in CDMA-based networks, referring to Section 6.1 for an overview of the relevant literature on packet scheduling evaluations.

The joint use of Transmission Power Control and rate selection to maximise the system throughput is considered in [129], allowing only a discrete set of feasible transfer rates, in contrast to the continuous range assumed in most papers. The integration of Transmission Power Control, variable forward error correction coding (link adaptation) and scheduling in the downlink of CDMA-based data networks is studied in [158], concluding that data scheduling based on link adaptation outperforms that based on Transmission Power Control in terms of the achieved system 'utility'.

Studies that concentrate on (some kind of) adaptive rate control in integrated services networks are rare. In [23], a heuristic load control scheme (de)activates the dedicated transport channels conveying data traffic in support of the varying speech traffic load. [202] investigates the effectiveness of decreasing data transfer rates to reduce interference levels as a form of congestion control. [88, 185] include a simulation study of a heuristic rate control scheme in an integrated speech/data scenario. While the cited references concentrate on heuristic and rather myopic schemes, the adaptive scheduling schemes considered in this chapter aim to *optimise* performance in a *network-wide* sense.

7.2. CONTRIBUTION

The presented study assesses the potential of fair and efficient adaptive power-based scheduling of data transfers in an integrated services UMTS network, serving speech and data calls. In view of the generally anticipated data traffic asymmetry, the downlink is expected to be the capacity bottleneck and therefore the focus of our attention. While the speech telephony service is mapped to DCHs, the downlink data transfers are scheduled on High-Speed Downlink Shared CHannels (HS-DSCHs, one per NodeB). The speech and data performance achieved by fixed and adaptive scheduling schemes is evaluated, while for each alternative two variants are specified according to distinct fairness objectives: one that distributes resources (transmission power) fairly over the data calls, and one that provides an equal throughput to all data calls. Analytical optimisation techniques are combined with Monte Carlo simulations. Numerical results are presented to demonstrate the potential performance enhancement adaptive scheduling can achieve for both service types, as well as to obtain insight in the performance implications of the distinct fairness objectives (*power*- versus *rate*-oriented).

7.3. HIGH-SPEED DOWNLINK PACKET ACCESS

A number of technological improvements of the initial UMTS system release are standardised under the name High-Speed Downlink Packet Access [2, 105, 157, 184, 185]. Equivalent, at least to some extent, to the EDGE [86, 87] evolution of second-generation (e.g. GSM) networks, the main objective of HSDPA in UMTS networks is to enable the support of downlink peak rates in the range of $8-10$ Mbits/s for best effort packet data services, i.e. far beyond the 3G requirement of 2 Mbits/s. To this end, HSDPA introduces the HS-DSCH as an upgraded version of the DSCH, which supports a variety of enhanced technologies, including:

HIGHER ORDER MODULATION: In addition to the QPSK (Quadrature Phase Shift Keying) modulation scheme specified in the 'basic' UMTS radio interface standards, the higher order 16-QAM (Quadrature Amplitude Modulation) modulation scheme is added in the HSDPA upgrades, in order to enhance spectral efficiency and thus enable higher data rates in favourable propagation and interference conditions. Since higher-order modulation is less robust to channel impairments, it should be combined with fast link adaptation.

FAST LINK ADAPTATION: Adaptive modulation and channel coding is applied at a small time scale based on terminal feedback in order to optimise data rates for actual channel conditions, e.g. higher-order modulation with little forward error correction redundancy for a terminal near its serving NodeB experiencing favourable fading conditions.

FAST SCHEDULING: A smaller transmission time interval of 2 ms (as opposed to the current 10 ms transmission time interval), is proposed as the heartbeat for the HS-DSCH allocation to different data flows, in order to reduce delays, allow a finer granularity of the scheduling process and facilitate better tracking of the channel variations. An extreme incidence of exploiting channel variations which greedily maximises instantaneous system throughput (but not necessarily the long-term average system throughput [33]) is pure C/I-based scheduling, i.e. always serve the terminal with the most favourable instantaneous channel conditions. There is an apparent trade-off between resource efficiency and fairness among data flows.

FAST CELL SELECTION: A fast cell selection procedure is deployed in order for a UE to be continuously served by the NodeB at which it experiences the best radio conditions, which is important in view of the unavailability of macro-diversity on the HS-DSCH. Fast cell selection can be regarded as the spatial equivalent to the temporal dimension of fast scheduling.

Other enhancements are an advanced hybrid ARQ scheme and the application of MIMO technology. As the effectiveness of the proposed technologies strongly relies on rapid adaptation of transmission parameters to the time-varying channel conditions, the corresponding control schemes, e.g. fast link adaptation and fast scheduling are best placed at the direct edge of the radio interface, i.e. at the NodeB. This is in contrast to the current UMTS architecture, where e.g. the scheduling function resides in the RNC.

The viability of fast link adaptation opens up a new spectrum of downlink scheduling options, since it now becomes possible to parameterise an HS-DSCH by an assigned transmission power (*power-based scheduling*[1]) [184], where the experienced data rate follows from the radio link quality of the served data call. As an alternative, *rate-based scheduling*[2], whose efficiency relies a.o. on the effectiveness of closed-loop Transmission Power Control rather than fast link adaptation, is deployed in basic UMTS networks (without the HSDPA upgrades). An important advantage of power- over rate-based scheduling is the reduced variability of the exerted power levels [13], since the path gain differences of the served data calls can be dealt with in the time domain: serving remote data calls a bit longer, rather than with greater transmission power. A remote data call thus influences other calls in the network in a smoother fashion by requiring a longer service at constant transmission power compared to near data calls, rather than requiring extreme powers when served by the scheduler. As a consequence of the reduced transmission power variability the Transmission Power Control performance of e.g. speech calls is expected to improve.

7.4. MODEL

The model description is broken up into four distinct segments concentrating on the system model, propagation model, traffic model and relevant performance measures.

[1] In the literature, *power-based* scheduling is also referred to as *rate-controlled* scheduling, where rate control refers to the *packet level* mechanism of selecting the optimal data rate depending on actual channel conditions (∼ link adaptation). The term 'power-based scheduling' is applied here in order to avoid confusion with the *call level* rate control mechanism as first introduced in Section 1.4.2.

[2] *Rate-based* scheduling is also referred to as *power-controlled* scheduling, where for a given transport channel rate the applied transmission power is adapted to the actual channel quality of the served data flow.

7.4.1. SYSTEM MODEL

We consider a cellular network of omnidirectional UMTS NodeBs in a hexagonal layout. In order to mimic an infinite network and thus avoid undesirable network boundary effects, the $B = 19$ reference cells are replicated as indicated in Figure 7.1. A hexagonal cell radius of $1/\sqrt{3} \approx 0.577$ km is assumed so that the inter-NodeB distance is precisely 1 km.

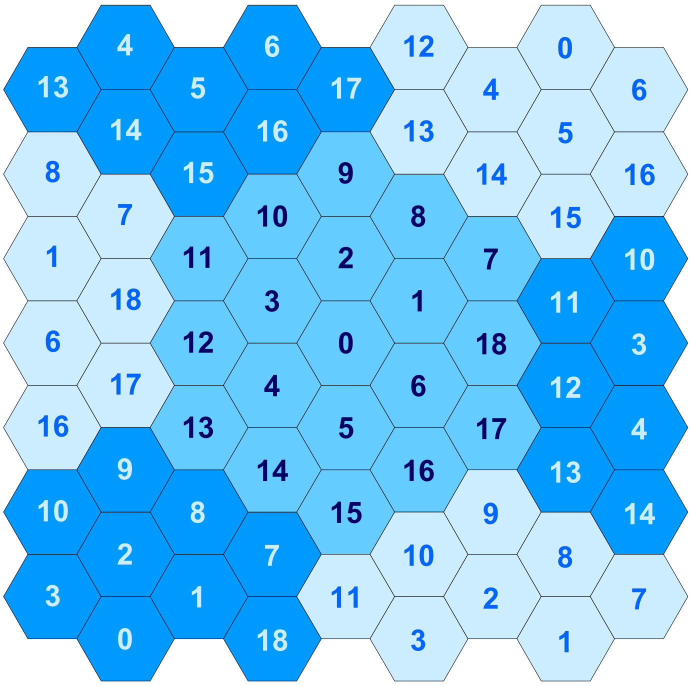


Figure 7.1 The considered UMTS network layout consists of a basic module of 19 omnidirectional NodeBs, which is replicated to mimic an infinite network.

As before, the presented investigation concentrates on the UMTS downlink in light of the anticipated traffic asymmetry. A NodeB's (downlink) transmission power budget is denoted $p_{\max}$ which is to be shared by the different services. The power consumption of common channels (e.g. paging, broadcast and forward access channels) is ignored. The spatially uniform thermal noise level is denoted $\nu > 0$.

7.4.2. PROPAGATION MODEL

The radio propagation model considers a single signal path with correlated lognormal shadowing. Given a distance r between the transmitter NodeB and the receiver UE, the relation between transmission ($p_{\text{transmission}}$) and reception ($p_{\text{reception}}$) power (in Watt) is given by

$$\begin{aligned} p_{\text{reception}} &= p_{\text{transmission}} \cdot \mathcal{G}_{\text{UE NodeB}} \\ &= p_{\text{transmission}} \cdot \eta_{\text{basic}} \cdot r^{-\varsigma} \cdot 10^{(a\xi_{\text{UE}}+b\xi_{\text{NodeB}})/10}, \end{aligned}$$

where η_{basic} reflects the basic transmission loss, ς is the attenuation exponent, $a\xi_{\text{UE}} + b\xi_{\text{NodeB}}$ is the correlated shadowing effect (in dB), with ξ_{UE}, $\xi_{\text{NodeB}} \sim N\left(0, \sigma_{\text{S}}^2\right)$ the mutually independent UE- and NodeB-specific shadowing effects and a and b the correlation factors [218]. $\mathcal{G}_{\text{UE NodeB}}$ denotes the local average path gain between the considered UE and NodeB, which is used to assign a serving NodeB to a given UE. As in Chapter 6, in C/I calculations ω denotes the downlink orthogonality factor, which applies to both time-shifted versions of identical ('self-orthogonality') or different ('cross-orthogonality') signals from the same NodeB. The concrete parameter settings of the propagation model are specified in Table 7.2 at the beginning of Section 7.7.

7.4.3. TRAFFIC MODEL

The considered UMTS network integrates speech and data services, whose relevant characteristics are summarised below. As the Monte Carlo-based evaluations concentrate on the performance gain that can be achieved by adaptive scheduling in a (randomly determined) snapshot of the UMTS network, given by the *number* of active speech and buffered data jobs, speech call durations or data call durations/sizes need not be specified.

SPEECH SERVICE: Speech calls are handled on Dedicated transport CHannels. The physical layer QOS target for speech calls is defined by an energy-per-bit to interference-plus-noise density ratio (E_b/N_o) target of $\gamma_{\text{speech}} = 7$ dB. Along with an information bit rate of $R_{\text{speech}} = 12.2$ kbits/s and a (fixed) system chip rate of $R_c = 3.84$ Mchips/s, this translates into a carrier-to-interference ratio (C/I) target of $\widetilde{\gamma}_{\text{speech}} \equiv (R_{\text{speech}}/R_c)\, 10^{0.5} \approx 0.0100$. It is assumed that

$\widetilde{\gamma}_{\text{speech}}^{-1} > \omega$ in order to allow at least a solitary speech call to be able to meet its C/I requirement despite the impaired self-orthogonality. No macro-diversity is included to allow the analytical treatment presented in Section 7.6, without violating our qualitative objectives.

DATA SERVICE: Data calls are assumed to be downlink data transfers which are handled on HS-DSCHs. One HS-DSCH is deployed per NodeB to multiplex the arriving data flows. Although in a real network each data call that is served on the HS-DSCH maintains a low bit rate associated DCH for control signalling purposes, this issue is ignored for transparency of the analysis and in line with our qualitative objectives. See Chapter 6 for a study on the impact of the associated DCHs on the data performance. The physical layer QOS target for data calls is defined by an E_b/N_o target of $\gamma_{\text{data}} = 4$ dB, which along with the (fixed) system chip rate R_c specifies a simple relation between the transfer rate R and the corresponding C/I target $\widetilde{\gamma}_{\text{data}}$:

$$R\gamma_{\text{data}} = R_c\widetilde{\gamma}_{\text{data}}, \tag{7.1}$$

(γ_{data} and $\widetilde{\gamma}_{\text{data}}$ in linear units). It is assumed that the implicitly assumed link adaptation feature (HSDPA) operates instantaneously and with infinitely small granularity, which is a reasonable assumption due to the qualitative nature of the formulated objectives. As we will see, relation (7.1) plays a key role in the design and evaluation of a variety of scheduling schemes.

The location of calls of either service type is uniformly sampled over the network, i.e. there are no structural hot or cold spots.

7.4.4. PERFORMANCE MEASURES

The performance of the *speech service* is expressed in terms of the outage probability, i.e. the probability that an arbitrary speech call fails to meet its C/I target. The performance of the *data service* is expressed in terms of the expected throughput per data call, i.e. the number of bits that are successfully transferred per second within a scheduling period. Additionally, we assess the speech and data performance *conditional* on the calls' geometries, a measure for terminal location, which incorporates the effects of shadowing.

7.5. SCHEDULING SCHEMES

This section proposes a set of power-based scheduling schemes which are specified as combinations of rate control and packet scheduling schemes, all providing some sort of *fairness* among data calls. A distinction is made between fairness from the network perspective, established by equalising the expected transmission power per data call, i.e. in terms of the assigned resources (*'power fairness'*), and fairness from the customer perspective, provided by equalising the expected transfer rate per data call, i.e. in terms of experienced throughputs (*'rate fairness'*).

At this point it is convenient to introduce the concept of *scheduling periods*, which establishes the heartbeat of the scheduling schemes' adaptivity. Refer to Figure 7.2 for an illustration of the different scheduling time scales. In practice it is advocated to reschedule HS-DSCH assignments (i.e. re-determine the rate control and packet scheduling parameters) whenever significant network state changes occur, e.g. at call arrival and termination events. As a consequence, such (largely unpredictable) events partition time in contiguous sequences of a variable number of time frames, called scheduling periods, during which the system is approximately constant. Within an adaptive scheduling period, the present data flows are multiplexed according to their current packet scheduling weights at the fixed heartbeat determined by the HSDPA transmission time intervals.

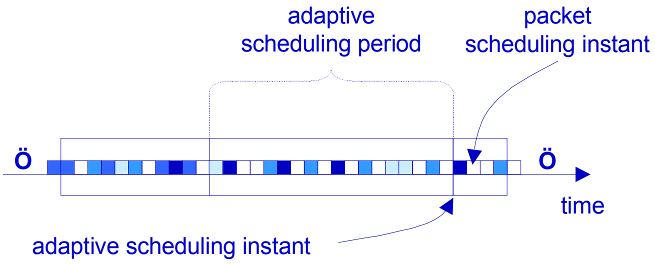


Figure 7.2 The investigated adaptive scheduling schemes consist of a rate control and a packet scheduling component with different heartbeats.

The investigated scheduling schemes are characterised by distinct implementations at the rate control and packet scheduling levels.

RATE CONTROL: The rate control schemes are in charge of assigning HS-DSCH transmission powers to the different NodeBs for the duration of the scheduling period in a *fixed* or *adaptive* manner. As the most likely default option, rate control schemes of the former type assign a fixed homogeneous HS-DSCH transmission power to all (non-idle) NodeBs, and thus operate independently of actual speech traffic load and the deployed packet scheduling scheme. A sort of spatial fairness is established in the sense that the network's resources are distributed evenly over the coverage area.

In contrast, *adaptive* rate control schemes aim to optimise data performance by maximising the resources assigned to the HS-DSCHs, under the condition that all speech calls meet their C/I requirements. Hence speech calls do not 'see' the data calls so that the speech performance is insensitive to the precise specifications of the adaptive scheduling scheme. Such rate control schemes operate in conjunction with the packet scheduling schemes in order to assign either equal expected transmission *powers* or transfer *rates* to all data calls. Both alternatives typically establish a heterogeneous power distribution over the NodeBs, determined such that at each NodeB the associated packet scheduling scheme is able to distribute the assigned resources in compliance with the posed objective of power or rate fairness. The adaptive rate control schemes typically shift resources from data traffic (incidentally) cold to (incidentally) hot spots in an attempt to enhance both throughputs and fairness.

PACKET SCHEDULING: Within each cell, the HS-DSCH transmission power assigned by the rate control scheme is distributed over the present data calls by the deployed packet scheduling scheme, which is in charge of time-multiplexing the different data transfers over the single HS-DSCH. Given our fairness objectives, two distinct packet scheduling flavours are considered. Either pure Round Robin scheduling is used to provide a fair intracellular allocation of the available transmission *power*, or a Weighted Round Robin (WRR) scheme is selected, with weights appropriately specified to establish that each data call experiences the same transfer *rate*, assigning greater scheduling weights to more remote terminals.

It is stressed that the latter (rate-fair) scheme is importantly different from simple RR scheduling on a *rate*-based (power-controlled) DSCH (as opposed to the *power*-based HS-DSCH considered here) as studied in Chapter 6, since the interference generated by a DSCH varies on the packet scheduling time scale of milliseconds, while a HS-DSCH changes its transmission power only at the larger time scale of rate control instants.

Table 7.1 summarises the categorisation of schemes, specifying four distinct combinations of rate control and packet scheduling schemes. The FP scheme combines Fixed rate control with RR (equal Powers within each cell) packet scheduling, while the FR scheme combines Fixed rate control with WRR (equal transfer Rates within each cell) packet scheduling. The AP scheme provides network-wide Power fairness by appropriately integrating Adaptive rate control with RR packet scheduling. Finally, network-wide Rate fairness is established by the AR scheme using Adaptive rate control in conjunction with WRR packet scheduling.

Table 7.1 Overview of scheduling schemes. A distinction is made between fixed and adaptive scheduling schemes on the one hand, and the different fairness objectives on the other hand.

	POWER FAIRNESS	RATE FAIRNESS
FIXED SCHEDULING	FP	FR
ADAPTIVE SCHEDULING	AP	AR

7.6. PERFORMANCE ANALYSIS

The relative performance of the presented scheduling schemes is evaluated by means of Monte Carlo simulations. Below we first introduce the necessary notation to specify a Monte Carlo snapshot, referred to as a *constellation*, which specifies the system state at a given scheduling instant, define the characterising parameters that specify the different scheduling schemes, and formulate how the constellation-specific performance is aggregated into the desired performance measures. Subsequently, for each of the considered scheduling schemes, the performance evaluation procedure is given to determine the constellation-specific performance.

CONSTELLATIONS

In order to evaluate the relative performance of the different scheduling schemes, K independent constellations are generated, denoted $(\mathbb{M}, \mathcal{G}, \mathbf{b})^k$, $k \in \mathbb{K} \equiv \{1, \cdots, K\}$, in order to take into account the random effects of terminal locations and shadowing.

A constellation is specified as follows. Speech and data calls are placed in the 19-cell wraparound network according to spatial Poisson processes with means ρ_{speech} and ρ_{data}, respectively. Denote by M^k_{speech} and M^k_{data} the sampled number of speech and data calls. Let $\mathbb{M}^k_{\text{speech}} \equiv \left\{1, \cdots, M^k_{\text{speech}}\right\}$, $\mathbb{M}^k_{\text{data}} \equiv \left\{1, \cdots, M^k_{\text{data}}\right\}$ and $\mathbb{B} \equiv \{1, \cdots, B\}$, denote the set of speech calls, data calls and NodeBs, respectively, and let $\mathbb{M}^k \equiv \mathbb{M}^k_{\text{speech}} \cup \mathbb{M}^k_{\text{data}}$. For each constellation $k \in \mathbb{K}$, the path gains $\mathcal{G}^k_{mb}$, $m \in \mathbb{M}^k$, $b \in \mathbb{B}$, are sampled and the UEs' NodeB assignments $\mathbf{b}^k \equiv \left(b^k_m \in \mathbb{B},\ m \in \mathbb{M}^k\right)$ are determined based on the 'maximum path gain criterion', i.e. each UE is assigned to the NodeB that provides the strongest radio link. Let $\mathbb{M}^{k,b}_{\text{speech}} \equiv \left\{m \in \mathbb{M}^k_{\text{speech}} : b_m = b\right\}$ ($M^{k,b}_{\text{speech}} \equiv \left|\mathbb{M}^{k,b}_{\text{speech}}\right|$) and $\mathbb{M}^{k,b}_{\text{data}} \equiv \left\{m \in \mathbb{M}^k_{\text{data}} : b_m = b\right\}$ ($M^{k,b}_{\text{data}} \equiv \left|\mathbb{M}^{k,b}_{\text{data}}\right|$) denote the set (number) of speech and data calls served by NodeB $b \in \mathbb{B}$, respectively. Hence $\mathbb{B}^k_+ \equiv \left\{b \in \mathbb{B} : \mathbb{M}^{k,b}_{\text{data}} \neq \emptyset\right\}$ ($B^k_+ \equiv \left|\mathbb{B}^k_+\right|$) is the set (number) of NodeBs that serves at least one data call in constellation k. A constellation $(\mathbb{M}, \mathcal{G}, \mathbf{b})^k$ is inherently (approximately) constant for the duration of the scheduling period (see Section 7.5).

CHARACTERISING PARAMETERS OF SCHEDULING SCHEMES

Some additional notation is required in order to specify the scheduling schemes in detail. If $P_{\text{HS-DSCH}}$ denotes the aggregate transmission power that the B NodeBs assign to the HS-DSCHs, the rate control weights φ_b denote the fraction of $P_{\text{HS-DSCH}}$ that is consumed by NodeB $b \in \mathbb{B}$. At the packet scheduling level, the HS-DSCH transmission power $\varphi_b P_{\text{HS-DSCH}}$ that is assigned to NodeB $b \in \mathbb{B}$ is allocated to data call $m \in \mathbb{M}^{k,b}_{\text{data}}$ during a fraction ϕ^b_m of the scheduling period, with $\sum_{m \in \mathbb{M}^{k,b}_{\text{data}}} \phi^b_m = 1$.

PERFORMANCE MEASURES

The speech performance achieved in constellation $k \in \mathbb{K}$ is measured in terms of the outage events, i.e. the event that speech call m does not meet its C/I target,

$$O^k_m \equiv \begin{cases} 1 & \text{if call } m \text{ in constellation } k \text{ experiences an outage} \\ \\ 0 & \text{otherwise} \end{cases}$$

$m \in \mathbb{M}^k_{\text{speech}}$. The data performance is expressed in terms of the achieved bit rate of data call m in constellation $k \in \mathbb{K}$, denoted with R^k_m, $m \in \mathbb{M}^k_{\text{data}}$. In order to evaluate the impact of terminal location on the experienced performance, the speech and data performance are also determined *conditional* on the geometry Ω (see also Section 6.7.2).

When all K constellations have been processed, the overall speech and data performance are expressed in terms of the *speech call outage probability* and the *expected throughput per data call*, given by

$$\mathbf{P}_{\text{speech}} \equiv \frac{\sum\limits_{k\in\mathbb{K}} \sum\limits_{m\in\mathbb{M}^k_{\text{speech}}} O^k_m}{\sum\limits_{k\in\mathbb{K}} M^k_{\text{speech}}} \quad \text{and} \quad \mathbf{R}_{\text{data}} \equiv \frac{\sum\limits_{k\in\mathbb{K}} \sum\limits_{m\in\mathbb{M}^k_{\text{data}}} R^k_m}{\sum\limits_{k\in\mathbb{K}} M^k_{\text{data}}},$$

respectively. In addition, the *conditional speech call outage probability* $\mathbf{P}_{\text{speech}}(\Omega)$ and the *conditional expected throughput per data call* $\mathbf{R}_{\text{data}}(\Omega)$ are determined.

The procedure to determine the O^k_m, $m \in \mathbb{M}^k_{\text{speech}}$, $k \in \mathbb{K}$, and R^k_m, $m \in \mathbb{M}^k_{\text{data}}$, $k \in \mathbb{K}$, for each of the considered scheduling schemes is outlined below. For enhanced readability, the index k indicating the considered constellation will be omitted in the remainder of this section.

7.6.1. FIXED SCHEMES

Under both the FP and the FR reference schemes each NodeB assigns a fixed transmission power of

$$p_b = p^{\star}_{\text{HS-DSCH}} 1 \left\{ \mathbb{M}^b_{\text{data}} \neq \emptyset \right\}, \; b \in \mathbb{B},$$

to its HS-DSCH, where $p^{\star}_{\text{HS-DSCH}}$ is fixed by the network planner. Observe that under the fixed schemes, which are considered for reference purposes, the HS-DSCH power assignments p_b, $b \in \mathbb{B}$, are fixed *directly* per NodeB, rather than first optimising the aggregate HS-DSCH transmission power $P_{\text{HS-DSCH}}$ and subsequently distributing this aggregate power over the non-idle NodeB's by means of the rate control weights φ_b, $b \in \mathbb{B}$ (see e.g. the AP scheme below). For completeness, we note that the above assignment implicitly specifies that

$$\varphi_b = B_+^{-1} 1 \left\{ \mathbb{M}_{\text{data}}^b \neq \emptyset \right\}, \ b \in \mathbb{B},$$

while the aggregate HS-DSCH transmission power is equal to $P_{\text{HS-DSCH}} = p_{\text{HS-DSCH}}^{\star} B_+$.

Since the (idealised) fast link adaptation scheme ensures that for any experienced C/I, a corresponding HS-DSCH bit rate can be chosen such that the associated C/I requirement is precisely equal to the experienced C/I, each constellation is inherently feasible with regard to the data service. The overall feasibility of a given constellation is then determined by whether the chosen power allocation for the HS-DSCHs allows the speech calls to achieve their C/I targets. Under a power-based scheduling scheme with HS-DSCH transmission powers $\mathbf{p}_{\text{data}} \equiv (p_b, \ b \in \mathbb{B})$, we call the constellation $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ *feasible* if and only if an assignment $\mathbf{p}_{\text{speech}} \equiv (p_m, \ m \in \mathbb{M}_{\text{speech}})$ of speech transmission powers exists, such that the C/I requirements of all speech calls are met:

$$(\mathbb{M}, \mathcal{G}, \mathbf{b}) \text{ feasible} \quad \Leftrightarrow \quad \exists \mathbf{p}_{\text{speech}} \in \mathbb{R}_+^{M_{\text{speech}}} \text{ such that}$$

$$\begin{cases} \dfrac{p_m \mathcal{G}_{m b_m}}{I_m^{\text{speech}}\left(\mathbf{p}_{\text{speech}}\right) + I_m^{\text{data}}\left(\mathbf{p}_{\text{data}}\right) + \nu} \geq \widetilde{\gamma}_{\text{speech}} & m \in \mathbb{M}_{\text{speech}} \\ \\ \displaystyle\sum_{m \in \mathbb{M}_{\text{speech}}^b} p_m \leq p_{\max} - p_b & b \in \mathbb{B} \end{cases}$$

where the different conditions reflect the speech calls' C/I requirements and the NodeBs' power budget limitations, the downlink orthogonality factor $\omega\left(b, b'\right)$ is as introduced in Section 6.6.1, and

$$I_m^{\text{speech}}\left(\mathbf{p}_{\text{speech}}\right) \equiv \sum_{m' \in \mathbb{M}_{\text{speech}}} \omega\left(b_m, b_{m'}\right) p_{m'} \mathcal{G}_{m b_{m'}},$$

and

$$I_m^{\text{data}}\left(\mathbf{p}_{\text{data}}\right) \equiv \sum_{b \in \mathbb{B}} \omega\left(b_m, b\right) p_b \mathcal{G}_{mb},$$

denote the amounts of interference call m experiences from speech and data transmissions, respectively, for given speech and transmission power vectors $\mathbf{p}_{\text{speech}}$ and $\mathbf{p}_{\text{data}}$.

The elaborations on the (in)feasibility of a constellation and on implementable distributed Transmission Power Control schemes that are able to achieve (sub)optimal power settings, given in the intermezzo 'ON CONSTELLATION (IN)FEASIBILITY' of Chapter 6, apply to the analogously specified constellations of this chapter as well. It is noted that *if* a constellation is feasible, then the C/I requirements can be satisfied, *with equality*, simultaneously on all radio links and hence none of the speech calls experiences an outage. The suggested Transmission Power Control scheme operates such that in an *infeasible* constellation, the transmission power of each call at a congested (power-limited) NodeB, is reduced proportionally, which corresponds with an approximately proportional reduction of the experienced C/I's. As a consequence, still some (not all) of the speech calls may meet their C/I targets, typically in lightly loaded cells.

In order to determine the performance for the data service, the distinction in packet scheduling schemes (FP versus FR) needs to be effectuated. Before doing so, we first note that since the data calls inherently meet their C/I targets via fast link adaptation, regardless of whether the obtained speech transmission power assignment $\mathbf{p}_{\text{speech}}$ satisfies the speech calls' C/I requirements, data call $m \in \mathbb{M}_{\text{data}}$ achieves a throughput of

$$R'_m = \frac{R_c}{\gamma_{\text{data}}} \frac{p^{\star}_{\text{HS-DSCH}} \mathcal{G}_{mb_m}}{I_m^{\text{speech}}\left(\mathbf{p}_{\text{speech}}\right) + I_m^{\text{data}}\left(\mathbf{p}_{\text{data}}\right) + \nu}$$

kbits/s whenever it is served (using (7.1)), so that the expected throughput during the scheduling period is given by

$$R_m = \phi_m^{b_m} R'_m,$$

where the packet scheduling weights $\phi_m^{b_m}$ depend on the deployed packet scheduling alternative.

Under the FP scheme, which operates a Round Robin packet scheduling mechanism, data call $m \in \mathbb{M}_{\text{data}}$ is assigned the HS-DSCH transmission power $p^{\star}_{\text{HS-DSCH}}$ at its serving NodeB b_m during a fraction

$$\phi_m^{b_m} = \left(M_{\text{data}}^{b_m}\right)^{-1}$$

of the scheduling period. Alternatively, under the FR scheme, which operates a Weighted Round Robin packet scheduling mechanism, data call $m \in \mathbb{M}_{\text{data}}$ is served during a fraction

$$\phi_m^{b_m} = \left(R'_m\right)^{-1} \left[\sum_{m' \in \mathbb{M}_{\text{data}}^{b_m}} \left(R'_{m'}\right)^{-1} \right]^{-1}$$

of the scheduling period. It is readily verified that these $\phi_m^{b_m}$'s indeed sum up to 1 at each NodeB, and that the established R_m's are indeed identical for all UEs at any given NodeB, as intended.

7.6.2. ADAPTIVE SCHEME WITH POWER FAIRNESS

The objective of the AP scheduling scheme is to provide network-wide power fairness, by assigning a uniform expected transmission power (an equal amount of resources) to all data calls. The rate control scheme assigns a transmission power of

$$p_b = \varphi_b P_{\text{HS-DSCH}}, \text{ with } \varphi_b = \frac{M_{\text{data}}^b}{M_{\text{data}}}, \; b \in \mathbb{B},$$

to NodeB b's HS-DSCH, which is combined with a Round Robin packet level scheduling scheme. The aggregate HS-DSCH transmission power $P_{\text{HS-DSCH}}$ is an optimisation parameter within each constellation (scheduling period). The objective is to maximise $P_{\text{HS-DSCH}}$ and hence the data performance, subject to the condition that the speech calls satisfy their C/I requirements, given by

$$\frac{p_m \mathcal{G}_{mb_m}}{I_m^{\text{speech}}\left(\mathbf{p}_{\text{speech}}\right) + P_{\text{HS-DSCH}} I_m^{\text{data}}\left(\boldsymbol{\varphi}\right) + \nu} \geq \widetilde{\gamma}_{\text{speech}}, \; m \in \mathbb{M}_{\text{speech}},$$

with $I_m^{\text{speech}}(\cdot)$ and $I_m^{\text{data}}(\cdot)$ as specified in Section 7.6.1, where the latter expression now takes a rate control rather than a transmission power vector as an argument, and thus expresses the amount of interference call m experiences from data transmissions *per exerted Watt of* $P_{\text{HS-DSCH}}$, for a given rate control vector $\boldsymbol{\varphi} \equiv (\varphi_b,\ b \in \mathbb{B})$.

For a given value of $P_{\text{HS-DSCH}}$, the constellation's (in)feasibility regarding the speech calls can be determined as under the FP and FR schemes. In particular, if constellation $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ is *infeasible for* $P_{\text{HS-DSCH}} = 0$, i.e. if the speech calls' C/I targets cannot even be satisfied in the absence of data transfer, the speech outage performance is registered in accordance with the outcome of the applied TPC procedure, while all data calls are assigned a throughput of 0 kbits/s. Consider now the case that the constellation is *feasible for* $P_{\text{HS-DSCH}} = 0$, where all speech calls meet their C/I targets.

Proposition 7.1 *If constellation* $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ *is feasible for* $P_{\text{HS-DSCH}} = 0$, *the optimal value of* $P_{\text{HS-DSCH}}$ *is equal to*

$$P^{\star}_{\text{HS-DSCH}} \equiv \min_{b \in \mathbb{B}} \left\{ \frac{p_{\max} - \boldsymbol{\zeta}_b^T (\mathcal{I} - \mathcal{H})^{-1} \widehat{\boldsymbol{\nu}}}{\boldsymbol{\zeta}_b^T (\mathcal{I} - \mathcal{H})^{-1} \widehat{\mathbf{I}} + \varphi_b} \right\}, \tag{7.2}$$

where

$$\mathcal{H} \equiv \left(\frac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{speech}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \cdot 1\left\{m \neq m'\right\} \right)_{mm'}, \ \textit{for } m, m' \in \mathbb{M}_{speech},$$

$$\widehat{\mathbf{I}} \equiv \left(\frac{I_m^{\text{data}}(\boldsymbol{\varphi})}{\left(\widetilde{\gamma}_{speech}^{-1} - \omega\right) \mathcal{G}_{mb_m}},\ m \in \mathbb{M}_{speech} \right),$$

captures the interference due to data transmissions,

$$\widehat{\boldsymbol{\nu}} \equiv \left(\frac{\nu}{\left(\widetilde{\gamma}_{speech}^{-1} - \omega\right) \mathcal{G}_{mb_m}},\ m \in \mathbb{M}_{speech} \right),$$

captures the thermal noise effects, and

$$\boldsymbol{\zeta}_b \equiv \left(1\left\{b_m = b\right\},\ m \in \mathbb{M}_{speech}\right),$$

indicates which speech calls are served by NodeB $b \in \mathbb{B}$ *(the superscript '*T*' indicates the transposition of a vector).*

Proof Along similar lines as followed in the intermezzo 'ON CONSTELLATION (IN)FEASIBILITY' in Chapter 6, the speech calls' C/I conditions for a given value of $P_{\text{HS-DSCH}}$ can be written in matrix form:

$$(\mathcal{I} - \mathcal{H})\,\mathbf{p}_{\text{speech}} \geq P_{\text{HS-DSCH}}\widehat{\mathbf{I}} + \widehat{\boldsymbol{\nu}}.$$

Momentarily ignoring the restrictions posed by the NodeB power budget $p_{\max}$, the Pareto optimal vector of speech transmission powers is given by (6.5):

$$\mathbf{p}^{\star}_{\text{speech}} \equiv (\mathcal{I} - \mathcal{H})^{-1}\left\{P_{\text{HS-DSCH}}\widehat{\mathbf{I}} + \widehat{\boldsymbol{\nu}}\right\},$$

which is non-negative and increasing in $P_{\text{HS-DSCH}}$, since the matrix $(\mathcal{I} - \mathcal{H})^{-1}$ is positive component-wise (in the considered case that the constellation is feasible for $P_{\text{HS-DSCH}} = 0$) [12]. Note that the above expression for $\mathbf{p}^{\star}_{\text{speech}}$ indicates that $P_{\text{HS-DSCH}}\widehat{\mathbf{I}}$ can be regarded as the vector of noise rises experienced by the active speech calls due to the assigned HS-DSCH transmission powers $P_{\text{HS-DSCH}}$. A necessary and sufficient restriction for $P_{\text{HS-DSCH}}$ from NodeB b's perspective is given by

$$\boldsymbol{\zeta}_b^{\mathcal{T}}\mathbf{p}^{\star}_{\text{speech}} + \varphi_b P_{\text{HS-DSCH}} \leq p_{\max} \Leftrightarrow P_{\text{HS-DSCH}} \leq \frac{p_{\max} - \boldsymbol{\zeta}_b^{\mathrm{T}}(\mathcal{I} - \mathcal{H})^{-1}\widehat{\boldsymbol{\nu}}}{\boldsymbol{\zeta}_b^{\mathrm{T}}(\mathcal{I} - \mathcal{H})^{-1}\widehat{\mathbf{I}} + \varphi_b},$$

for each $b \in \mathbb{B}$. Since the restrictions must hold for all NodeBs, (7.2) follows. □

Remark 7.1 In expression (7.2) the numerator $p_{\max} - \boldsymbol{\zeta}_b^{\mathcal{T}}(\mathcal{I} - \mathcal{H})^{-1}\widehat{\boldsymbol{\nu}}$ for NodeB $b \in \mathbf{B}$ is equal to the NodeB's unused transmission power in the case where $P_{\text{HS-DSCH}} = 0$, which is non-negative due to the assumed feasibility of the constellation. The denominator $\boldsymbol{\zeta}_b^{\mathcal{T}}(\mathcal{I} - \mathcal{H})^{-1}\widehat{\mathbf{I}} + \varphi_b$ equals the amount of additional transmission power an additional Watt of $P_{\text{HS-DSCH}}$ requires at the NodeB, consisting of the additional power needed to satisfy the speech calls ($\boldsymbol{\zeta}_b^{\mathcal{T}}(\mathcal{I} - \mathcal{H})^{-1}\widehat{\mathbf{I}}$) and the HS-DSCH power itself (φ_b). Note that if NodeB b serves no speech calls ($\boldsymbol{\zeta}_b = \mathbf{0}$), the restriction imposed by this NodeB reduces to $\varphi_b P_{\text{HS-DSCH}} \leq p_{\max}$, while if it serves no data calls ($\varphi_b = 0$), the NodeB still imposes a restriction on $P_{\text{HS-DSCH}}$ due to the interference its speech calls experience from HS-DSCHs in other cells.

Remark 7.2 Proposition 7.1 captures the optimal HS-DSCH power level for a class of adaptive rate control schemes, that can be characterised using scheduling weights φ_b, $b \in \mathbb{B}$. For example, if a network operator wants to maintain the spatial fairness of the fixed schemes, yet aspires to optimise the *uniform* HS-DSCH transmission power assignment to all non-idle NodeBs, the above proposition can be readily applied using $\varphi_b = 1\left\{b \in \mathbb{B}_+\right\}/B_+$, $b \in \mathbb{B}$.

The throughput attained by data call $m \in \mathbb{M}_{\text{data}}$, when served, is given by

$$R'_m = \frac{R_c}{\gamma_{\text{data}}} \frac{\varphi_{b_m} P^{\star}_{\text{HS-DSCH}} \mathcal{G}_{mb_m}}{I^{\text{speech}}_m \left(\mathbf{p}^{\star}_{\text{speech}}\right) + P^{\star}_{\text{HS-DSCH}} I^{\text{data}}_m \left(\boldsymbol{\varphi}\right) + \nu}$$

kbits/s, so that the expected throughputs during the scheduling period are given by

$$R_m = \phi^{b_m}_m R'_m = R'_m / M^{b_m}_{\text{data}},$$

for $m \in \mathbb{M}_{\text{data}}$, with $\phi^{b_m}_m = 1/M^{b_m}_{\text{data}}$ the Round Robin packet scheduling weights.

7.6.3. ADAPTIVE SCHEME WITH RATE FAIRNESS

The objective of the AR scheduling scheme is to determine the appropriate combination of the aggregate HS-DSCH transmission power $P_{\text{HS-DSCH}}$, rate control (φ_b, $b \in \mathbb{B}$) and packet scheduling weights $\left(\phi^b_m,\ b \in \mathbb{B},\ m \in \mathbb{M}^b_{\text{data}}\right)$, which optimises the *uniform* expected throughput (rate fairness). This complex task requires an integrated approach to rate control and packet scheduling, where complete knowledge of all UE's path gain vectors is required at both scheduling levels. In contrast, recall that under the AP scheme, the specification of the rate control weights merely required the *number* of data calls in each cell, rather than their path gain information.

Denote with R the uniform expected throughput per data call, which is to be maximised under the condition that all speech calls meet their C/I targets. For $R = 0$, i.e. in the absence of data transfer, the constellation's (in)feasibility regarding the speech calls (outage events) corresponds to that under the AP scheme with $P_{\text{HS-DSCH}} = 0$. *Infeasibility* of constellation $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ for $R = 0$ implies zero throughput for all data calls, i.e. the optimal uniform expected throughput equals $R^{\star} = 0$. In case the constellation is *feasible for* $R = 0$, where no speech call experiences an outage, Proposition 7.2 gives the optimal uniform data throughput $R_m = R^\star$, $m \in \mathbb{M}_{\text{data}}$.

Proposition 7.2 *If constellation* $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ *is feasible for* $R = 0$*, the optimal value of* R *is equal to*

$$\begin{aligned} R^\star &\equiv \max R, \\ & \textit{subject to} \left\{ \begin{array}{l} \lambda_{\mathcal{H}} < 1 \\ \\ \zeta_b^{\mathcal{T}} \left(\mathcal{I} - \mathcal{H}\right)^{-1} \widehat{\boldsymbol{\nu}} \leq p_{\max} \quad b \in \mathbb{B} \end{array} \right. \end{aligned} \tag{7.3}$$

where the first condition reflects the C/I *requirements, and the second set of conditions is imposed by the* NodeBs*' power budget limitations. Here* $\lambda_{\mathcal{H}}$ *is the Perron-Frobenius eigenvalue of the non-negative* $(M_{speech} + B_+) \times (M_{speech} + B_+)$*-dimensional matrix* $\mathcal{H}$ *(see e.g.* [208, Chapter 1]*), partitioned as follows:*

$$\begin{aligned} \mathcal{H} &\equiv \begin{pmatrix} \mathcal{H}_{\mathbb{M}_{speech}\mathbb{M}_{speech}} & \mathcal{H}_{\mathbb{M}_{speech}\mathbb{B}_+} \\ \mathcal{H}_{\mathbb{B}_+\mathbb{M}_{speech}} & \mathcal{H}_{\mathbb{B}_+\mathbb{B}_+} \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{\omega\left(b_m, b_{m'}\right) \mathcal{G}_{mb_{m'}}}{\left(\widetilde{\gamma}_{speech}^{-1} - \omega\right) \mathcal{G}_{mb_m}} 1\left\{m \neq m'\right\} & \dfrac{\omega\left(b_m, b\right) \mathcal{G}_{mb}}{\left(\widetilde{\gamma}_{speech}^{-1} - \omega\right) \mathcal{G}_{mb_m}} \\ \\ \dfrac{\sum\limits_{m' \in \mathbb{M}_{\text{data}}^b} \omega\left(b, b_m\right) \frac{\mathcal{G}_{m'b_m}}{\mathcal{G}_{m'b}}}{\frac{R_c}{R\gamma_{\text{data}}} - M_{\text{data}}^b \omega} & \dfrac{\sum\limits_{m \in \mathbb{M}_{\text{data}}^b} \frac{\mathcal{G}_{mb'}}{\mathcal{G}_{mb}}}{\frac{R_c}{R\gamma_{\text{data}}} - M_{\text{data}}^b \omega} 1\left\{b \neq b'\right\} \end{pmatrix}, \end{aligned}$$

where $m, m' \in \mathbb{M}_{speech}$ *in* $\mathcal{H}_{\mathbb{M}_{speech}\mathbb{M}_{speech}}$*,* $m \in \mathbb{M}_{speech}$ *and* $b \in \mathbb{B}_+$ *in* $\mathcal{H}_{\mathbb{M}_{speech}\mathbb{B}_+}$*,* $b \in \mathbb{B}_+$ *and* $m \in \mathbb{M}_{speech}$ *in* $\mathcal{H}_{\mathbb{B}_+\mathbb{M}_{speech}}$*, and* $b, b' \in \mathbb{B}_+$ *in* $\mathcal{H}_{\mathbb{B}_+\mathbb{B}_+}$*. The* $(M_{speech} + B_+)$*-dimensional vector* $\widehat{\boldsymbol{\nu}}$ *is given by*

$$\begin{aligned} \widehat{\boldsymbol{\nu}} &\equiv \left(\widehat{\boldsymbol{\nu}}_{\mathbb{M}_{speech}}, \widehat{\boldsymbol{\nu}}_{\mathbb{B}_+}\right) \\ &= \left(\left(\frac{\nu}{\left(\widetilde{\gamma}_{speech}^{-1} - \omega\right) \mathcal{G}_{mb_m}},\ m \in \mathbb{M}_{speech}\right), \left(\frac{\nu \sum\limits_{m \in \mathbb{M}_{\text{data}}^b} \mathcal{G}_{mb}^{-1}}{\left(\frac{R_c}{R\gamma_{\text{data}}} - M_{\text{data}}^b \omega\right)},\ b \in \mathbb{B}_+\right)\right), \end{aligned}$$

and

$$\boldsymbol{\zeta}_b \equiv \left(\left(1\left\{b_m = b\right\},\ m \in \mathbb{M}_{speech}\right), \left(1\left\{b' = b\right\},\ b' \in \mathbb{B}_+\right)\right),$$

indicating a.o. which speech calls are served by NodeB $b \in \mathbb{B}_+$. *Note that both* $\mathcal{H}$ *and* $\widehat{\boldsymbol{\nu}}$ *depend on* R.

Proof For a given uniform transfer rate R the following necessary and sufficient conditions need to hold for constellation $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ to be feasible:

$(\mathbb{M}, \mathcal{G}, \mathbf{b})|_R$ feasible $\Leftrightarrow$

$\exists \mathbf{p}_{\text{speech}} \in \mathbb{R}_+^{M_{\text{speech}}}$, $\mathbf{p}_{\text{data}} \in \mathbb{R}_+^{B}$ and $\phi_m^b \in [0,1]$, $m \in \mathbb{M}_{\text{data}}^b$, $b \in \mathbb{B}$, such that

$$\left\{\begin{array}{lll} \dfrac{p_m \mathcal{G}_{mb_m}}{I_m^{\text{speech}}\left(\mathbf{p}_{\text{speech}}\right) + I_m^{\text{data}}\left(\mathbf{p}_{\text{data}}\right) + \nu} \geq \widetilde{\gamma}_{\text{speech}} & m \in \mathbb{M}_{\text{speech}}, \\ \dfrac{p_{b_m} \mathcal{G}_{mb_m}}{I_m^{\text{speech}}\left(\mathbf{p}_{\text{speech}}\right) + I_m^{\text{data}}\left(\mathbf{p}_{\text{data}}\right) + \nu} \geq \dfrac{R\gamma_{\text{data}}}{\phi_m^{b_m} R_c} & m \in \mathbb{M}_{\text{data}}, \\ \displaystyle\sum_{m \in \mathbb{M}_{\text{speech}}^b} p_m + p_b \leq p_{\max} & b \in \mathbb{B}, \\ \displaystyle\sum_{m \in \mathbb{M}_{\text{data}}^b} \phi_m^b = 1 & b \in \mathbb{B}_+, \\ p_b = 0 & b \in \mathbb{B} \backslash \mathbb{B}_+, \end{array}\right. \tag{7.4}$$

where the different sets of conditions reflect the C/I requirements of speech and data calls, the NodeBs' power budget limitations, the fact that the packet scheduling weights must sum up to 1, and the silencing of idle DSCHs. In the above expressions, $I_m^{\text{speech}}(\cdot)$ and $I_m^{\text{data}}(\cdot)$ are as specified in Section 7.6.1. This constitutes a set of $2B + M_{\text{speech}} + M_{\text{data}}$ conditions in $B + M_{\text{speech}} + M_{\text{data}}$ unknowns ($\mathbf{p}_{\text{speech}}$, $\mathbf{p}_{\text{data}}$ and $\left(\phi_m^{b_m},\ m \in \mathbb{M}_{\text{data}}\right)$).

A necessary feasibility condition for R that is readily verified to be implied by the data calls C/I requirements given above, follows from

$$\begin{aligned}\frac{p_b}{\omega p_b} &\geq \frac{R\gamma_{\text{data}}}{\phi_m^{b_m} R_c},\ \forall m \in \mathbb{M}_{\text{data}}^b \Rightarrow 1 = \sum_{m\in\mathbb{M}_{\text{data}}^b} \phi_m^b \geq \omega M_{\text{data}}^b \frac{R\gamma_{\text{data}}}{R_c} \\ &\Leftrightarrow\ R \leq \frac{R_c}{\omega M_{\text{data}}^b \gamma_{\text{data}}},\end{aligned} \tag{7.5}$$

for all $b \in \mathbb{B}_+$, since otherwise the C/I required to support UE m's target rate of R/ϕ_m^b, $m \in \mathbb{M}_{\text{data}}^b$, cannot be achieved even when disregarding the multiple access interference and noise. Condition (7.5) is useful to obtain an upper bound for the optimal uniform transfer rate $R^\star$ and hence a useful starting point in the search for $R^\star$. In the following, R is therefore assumed to satisfy (7.5) for all $b \in \mathbb{B}_+$.

Combining the second and fourth set of conditions of (7.4) in a similar manner,

$$\begin{aligned}1 &= \sum_{m\in\mathbb{M}_{\text{data}}^b} \phi_m^b \geq \frac{R\gamma_{\text{data}}}{p_b R_c} \sum_{m\in\mathbb{M}_{\text{data}}^b} \frac{I_m^{\text{speech}}(\mathbf{p}_{\text{speech}}) + I_m^{\text{data}}(\mathbf{p}_{\text{data}}) + \nu}{\mathcal{G}_{mb}} \\ &\Leftrightarrow \left(\sum_{m\in\mathbb{M}_{\text{data}}^b} \frac{\left(I_m^{\text{speech}}(\mathbf{p}_{\text{speech}}) + I_m^{\text{data}}(\mathbf{p}_{\text{data}}) + \nu\right)}{p_b \mathcal{G}_{mb}}\right)^{-1} \geq \frac{R\gamma_{\text{data}}}{R_c},\end{aligned} \tag{7.6}$$

$b \in \mathbb{B}_+$, allows us to reduce the system to one with $B + M_{\text{speech}}$ unknowns (and $2B + M_{\text{speech}}$ conditions):

$$\begin{cases} \dfrac{p_m \mathcal{G}_{mb_m}}{I_m^{\text{speech}}(\mathbf{p}_{\text{speech}}) + I_m^{\text{data}}(\mathbf{p}_{\text{data}}) + \nu} \geq \tilde{\gamma}_{\text{speech}} & m \in \mathbb{M}_{\text{speech}}, \\ \left(\displaystyle\sum_{m\in\mathbb{M}_{\text{data}}^b} \frac{\left(I_m^{\text{speech}}(\mathbf{p}_{\text{speech}}) + I_m^{\text{data}}(\mathbf{p}_{\text{data}}) + \nu\right)}{p_b \mathcal{G}_{mb}}\right)^{-1} \geq \dfrac{R\gamma_{\text{data}}}{R_c} & b \in \mathbb{B}_+, \\ \displaystyle\sum_{m\in\mathbb{M}_{\text{speech}}^b} p_m + p_b \leq p_{\max} & b \in \mathbb{B}, \\ p_b = 0 & b \in \mathbb{B}\backslash\mathbb{B}_+. \end{cases}$$

Note that the conditions that aggregate the data calls' C/I requirements per NodeB incorporate the data call specific C/I requirements in a 'harmonic' fashion, which reflects the fact that the packet scheduling weight is inversely proportional to the experienced C/I. Algebraic manipulations similar to those applied in the intermezzo 'ON CONSTELLATION (IN)FEASIBILITY' in Chapter 6, simplify the first two sets of inequalities to the matrix inequality

$$(\mathcal{I} - \mathcal{H})\,\mathbf{p}_+ \geq \widehat{\boldsymbol{\nu}}, \tag{7.7}$$

with $\mathbf{p}_+ \equiv (\mathbf{p}_{\text{speech}};\ p_b,\ b \in \mathbb{B}_+)$, $\mathcal{H}$ and $\widehat{\boldsymbol{\nu}}$ as specified in Proposition 7.2. Note that the denominators which appear in $\mathcal{H}$ and $\widehat{\boldsymbol{\nu}}$ are guaranteed to be positive, using a.o. expression (7.5). Furthermore, note that $\mathcal{H}$ has non-negative elements only. As a sanity check, note that for $R = 0$, $\mathcal{H}_{\mathbb{B}_+\mathbb{M}_{\text{speech}}} = \mathcal{H}_{\mathbb{B}_+\mathbb{B}_+} = \mathcal{O}$ and $\widehat{\boldsymbol{\nu}}_{\mathbb{B}_+} = \mathbf{0}$, which leads to the correct Pareto-optimal solution that $p_b = 0$, $b \in \mathbb{B}$.

The condition that a non-negative power assignment vector exists that satisfies (7.7) is equivalent to the condition that $\lambda_{\mathcal{H}} < 1$ in (7.3) (see e.g. [208, Chapter 2]), as well as to the condition that $(\mathcal{I} - \mathcal{H})^{-1}$ exists and is positive component-wise (see e.g. [12]). If these equivalent conditions are satisfied, then

$$\mathbf{p}_+^\star \equiv (\mathcal{I} - \mathcal{H})^{-1}\,\widehat{\boldsymbol{\nu}},$$

is the Pareto-optimal power assignment vector, in the sense that any other $\mathbf{p}_+$ satisfying (7.7) would require at least as much power on each radio link. Note that $\mathbf{p}_+^\star$ satisfies the C/I requirements with equality on all radio links simultaneously. The power budget condition is then equivalent with

$$\zeta_b^\tau \mathbf{p}_+^\star \leq p_{\max} \Leftrightarrow \zeta_b^\tau\,(\mathcal{I} - \mathcal{H})^{-1}\,\widehat{\boldsymbol{\nu}} \leq p_{\max},$$

$b \in \mathbb{B}$, as formulated in (7.3). To conclude, the conditions formulated in (7.3) are both necessary and sufficient for the feasibility of constellation $(\mathbb{M}, \mathcal{G}, \mathbf{b})$ given R. □

Since the data calls' C/I requirements are readily seen to become more stringent for greater R, a constellation that is feasible for a given R is also feasible for all $R' < R$. Conversely, a constellation which is *in*feasible for a given R, is also *in*feasible for all $R' > R$. Hence a bisection search procedure can be used effectively to optimise R, using necessary condition (7.5) to derive a practical starting point.

Upon convergence of the bisection search procedure, the optimal uniform transfer rate $R^\star$ is obtained, while the corresponding Pareto-optimal power assignment is given by $\left(\mathbf{p}^\star_{\text{speech}}; \mathbf{p}^\star_{\text{data}}\right)$, which merges $\mathbf{p}^\star_+ \equiv (\mathcal{I} - \mathcal{H})^{-1}\,\widehat{\boldsymbol{\nu}}$, containing the optimal speech powers and the HS-DSCH powers for NodeBs $b \in \mathbb{B}_+$, where $\mathcal{H}$ and $\widehat{\boldsymbol{\nu}}$ correspond with the optimal uniform transfer rate $R^\star$, and $p^\star_b = 0$, $b \in \mathbb{B}\backslash\mathbb{B}_+$. Although the rate control settings are completely specified by these optimal NodeB transmission powers, we note for completeness that the aggregate HS-DSCH transmission power and the optimal rate control weights are given by

$$P^\star_{\text{HS-DSCH}} \equiv \sum_{b \in \mathbb{B}} p^\star_b \quad \text{and} \quad \varphi^\star_b \equiv \frac{p^\star_b}{P^\star_{\text{HS-DSCH}}},$$

respectively.

Given the optimal power assignments, data call $m \in \mathbb{M}_{\text{data}}$ experiences an instantaneous throughput, when served, of

$$R'_m = \frac{R_c}{\gamma_{\text{data}}} \frac{p^\star_{b_m} \mathcal{G}_{m b_m}}{I^{\text{speech}}_m \left(\mathbf{p}^\star_{\text{speech}}\right) + I^{\text{data}}_m \left(\mathbf{p}^\star_{\text{data}}\right) + \nu}$$

kbits/s. The WRR packet scheduling weights need to be set as

$$\phi^{b_m}_m = \left(R'_m\right)^{-1} \left[\sum_{m' \in \mathbb{M}^{b_m}_{\text{data}}} \left(R'_{m'}\right)^{-1} \right]^{-1}, \; m \in \mathbb{M}_{\text{data}},$$

in order to establish network wide rate fairness:

$$\begin{aligned} R_m &= \phi^{b_m}_m R'_m = \left[\sum_{m' \in \mathbb{M}^{b_m}_{\text{data}}} \left(R'_{m'}\right)^{-1} \right]^{-1} \\ &= \left[\sum_{m' \in \mathbb{M}^{b_m}_{\text{data}}} \left(\frac{R_c}{\gamma_{\text{data}}} \frac{p^\star_{b_{m'}} \mathcal{G}_{m' b_{m'}}}{I^{\text{speech}}_{m'} \left(\mathbf{p}^\star_{\text{speech}}\right) + I^{\text{data}}_{m'} \left(\mathbf{p}^\star_{\text{data}}\right) + \nu} \right)^{-1} \right]^{-1} = R^*, \end{aligned}$$

for all $m \in \mathbb{M}_{\text{data}}$. The last equality holds since the Pareto-optimal power assignment vector $\left(\mathbf{p}^{\star}_{\text{speech}}; \mathbf{p}^{\star}_{\text{data}}\right)$ satisfies all C/I requirements with equality (cf. condition (7.6), $\forall b \in \mathbb{B}_{+}$).

7.7. NUMERICAL RESULTS

A number of numerical experiments are conducted in order to assess the potential performance enhancements provided by the power- and rate-fair adaptive scheduling schemes, using their respective fixed counterparts for reference purposes, as well to generate some valuable insight into the relative performance implied by the distinct objectives of power and rate fairness. The conducted experiments are based on Monte Carlo simulations that inherently assume *persistent* speech and data calls and thus disregard the dynamics of call initiations and completions. Sensible inclusion of these call level dynamics would require an inevitably extremely time-consuming combination of the presented analytical scheduling parameter optimisation with dynamic system-level simulations. As will be demonstrated, the applied Monte Carlo experiments not only enable us to evaluate the performance gains from the scheduling schemes' adaptivity, the obtained numerical results also present some revealing insights into e.g. the degree of spatial (un)fairness that is effectuated by the power- and rate-fair scheduling schemes, even if these insights do not establish a conclusive comparison.

Table 7.2 lists the settings of the relevant system and traffic parameters that are prefixed at typical values (the maximum transmission power and thermal noise level correspond to 42 dBm and −99.157 dBm, respectively). For the non-adaptive FP/FR scheduling schemes, the uniform HS-DSCH transmission power is set at $p^{\star}_{\text{HS-DSCH}} \in \{2, 3\}$ Watt. These reference HS-DSCH transmission powers have been selected in order to demonstrate that the adaptive schemes, while *always* improving speech quality (at the sample path level), can either *reduce* or *improve* the *expected* data throughput, depending on $p^{\star}_{\text{HS-DSCH}}$. The number of evaluated constellations is taken sufficiently large to ensure that the relative precision of the 95%-confidence intervals that have been constructed for unconditional performance measures, does not exceed 5%.

Figures 7.3, 7.4 and 7.5 depict the speech and data performance for an illustrative traffic mix of $\left(\rho_{\text{speech}}, \rho_{\text{data}}\right) = (800, 200)$.

Table 7.2 Settings of the system, propagation and traffic model parameters for the numerical evaluations.

SYSTEM MODEL			PROPAGATION MODEL			TRAFFIC MODEL		
B	19	NodeBs	η_{basic}	137.744	dB	R_{speech}	12.2	kbits/s
radius	$1/\sqrt{3}$	km	ς	3.523		γ_{speech}	7	dB
R_c	3840	kchips/s	a	$1/\sqrt{2}$		γ_{data}	4	dB
$p_{\max}$	15.849	Watt	b	$1/\sqrt{2}$				
ν	$1.214\ 10^{-13}$	Watt	σ_{S}	8	dB			
			ω	0.400				

7.7.1. SPEECH PERFORMANCE

In Figure 7.3 (left) the speech call outage probability is shown for the different scheduling schemes conditional on the calls' geometry, along with the geometry PDF, which is implicitly determined by the assumed PDFs of user location and correlated shadowing. Recall from Section 6.7.2 that a low geometry value reflects a UE that is typically located in the border region of adjacent cells, while a large geometry value corresponds to a UE located relatively close to its serving NodeB. As is intuitively obvious the outage probability decreases with the geometry: it is easier to satisfy a speech call's C/I requirement when it is located relatively close to a NodeB, since it experiences little interference from surrounding NodeBs. This trend holds regardless of the deployed scheduling scheme. When comparing the scheduling schemes, we observe that the outage probability curves coincide for both adaptive scheduling schemes (AP and AR), since under these schemes speech traffic does not 'see' data traffic in the sense that data is transferred only if this does not inhibit the satisfaction of the speech calls' C/I requirements.

Also, the fixed (non-adaptive) scheduling schemes (FP and FR) yield the same performance for a given $p^{\star}_{\text{HS-DSCH}}$, since the interference levels experienced by speech calls are influenced by the HS-DSCH transmission power (rate control), not by its distribution over the data calls (packet scheduling). As expected, for the fixed schemes the speech outage probabilities are worse than under the adaptive (protective) schemes, and are increasing in the assigned transmission power $p^{\star}_{\text{HS-DSCH}}$. Figure 7.3 (right) shows the unconditional speech performance (averaged over the geometry PDF) and is in obvious agreement with the trends observed for the conditional performance.

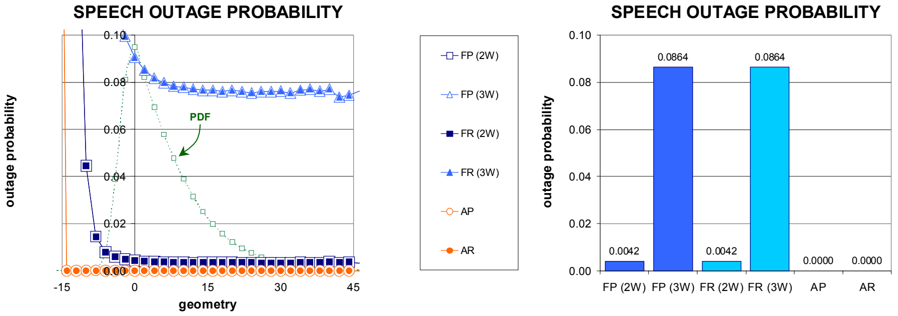


Figure 7.3 Speech performance for $(\rho_{\text{speech}}, \rho_{\text{data}}) = (800, 200)$. For different scheduling schemes, the left chart shows the conditional outage probability as a function of the speech calls' geometry, while the right chart depicts the unconditional speech performance.

7.7.2. DATA PERFORMANCE

Figure 7.4 depicts the geometry-conditional (left) and unconditional (right) data performance in terms of the expected throughputs (in kbits/s). Whereas the speech performance was fully determined by the rate control objectives (fixed versus adaptive), the data performance is clearly influenced by the packet scheduling operations as well. The left chart reveals the relative rate fairness that is established by the FR and AR schemes: the expected throughput is much less variant in the geometry than under the FP and AP schemes, which dramatically favour near over far data calls. Observe, however, that the *unconditional* expected data throughput is significantly higher under the power-fair schemes. An intuitive explanation of this interesting observation is readily given. Since the FP/AP schemes assign the HS-DSCH transmission powers regardless of the different data call path gains, the impact of a poor radio link on the data performance is limited to the corresponding data call only. In other words, although a badly located data call may suffer greatly from its bad link, it does not inflict its own suffering upon the other data calls and hence its unfavourable location leaves all other data calls' throughputs unaffected. In case of the FR/AR schemes, however, such a badly located data call requires a significant packet scheduling weight and thus allows only a low uniform data throughput to be achieved. Hence all data calls suffer from the poor radio link of a critical data call. Although such joint suffering may seem fair, the overall performance of the data service is reduced significantly, as demonstrated most clearly in Figure 7.4 (right). Observe, however, that the rate-fair (FR/AR) schemes do outperform the power-fair schemes (FP/AP) for remote data calls, even slightly beyond the mode of the geometry PDF.

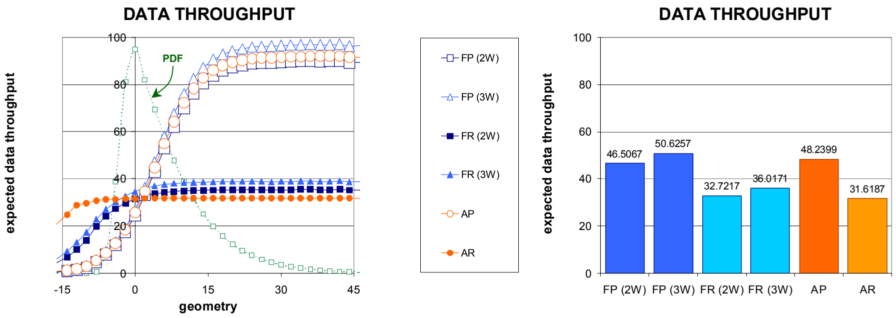


Figure 7.4 Data performance for $(\rho_{\text{speech}}, \rho_{\text{data}}) = (800, 200)$. For different scheduling schemes, the left chart shows the conditional data throughput as a function of the data calls' geometry, while the right chart depicts the unconditional data performance.

As a final observation, it is pointed out that the deployment of adaptive scheduling can indeed improve both the speech and data performance simultaneously. Compare e.g. the unconditional speech and data QOS experienced under the FP ($p^{\star}_{\text{HS-DSCH}} = 2$ Watt) and AP schemes (Figures 7.3 (right) and 7.4 (right)). Observe however that due to the strict priority given to the speech service, the data QOS gain is rather small. It is left open as a topic for further research to investigate the impact of a weaker speech service prioritisation scheme (e.g. a minimum resource reservation for data calls) on the speech and data QOS. While it is impossible for non-adaptive schemes to improve the speech performance unilaterally, simply setting an excessive HS-DSCH transmission power can increase the expected HS-DSCH throughputs unilaterally, but only at a dramatic degradation in speech performance.

Figure 7.5 (left) is included in order to demonstrate the power fairness that is established by the different FP and AP schemes, in the sense that the system's resources are fairly distributed over the data calls. While the expected transmission power per data call is obviously a perfectly flat curve for the FP/AP schemes, the curves corresponding to the FR/AR schemes reflect that providing uniform throughputs requires a higher (lower) transmission power assignment (effectively: a greater (smaller) WRR packet scheduling weight) to remote (near) data calls. It is interesting to observe that the curves of the power-fair and the corresponding rate-fair scheduling schemes cross at or close to the modal geometry, as was the case in Figure 7.4 (left). Apparently, the expected data performance experienced under each of the schemes is approximately equal for a typically located data call, while the differences between the schemes are most significantly experienced at less typical locations.

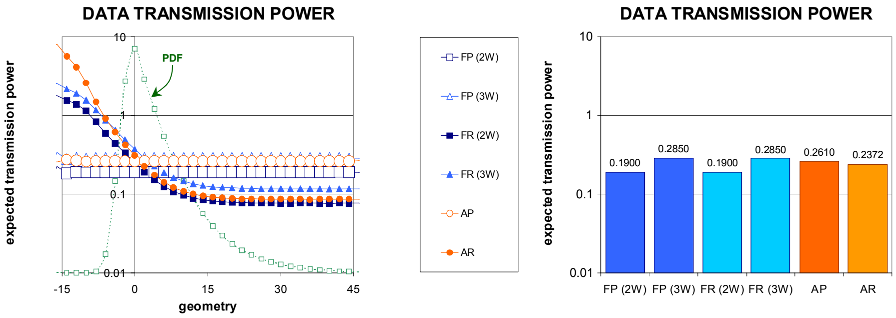


Figure 7.5 Allocated data transmission powers for $(\rho_{\text{speech}}, \rho_{\text{data}}) = (800, 200)$. For different scheduling schemes, the left chart shows the conditional assigned transmission power per data call as a function of the data calls' geometry, while the right chart depicts the unconditional transmission powers per data call.

The unconditional expected transmission power per data call is shown in Figure 7.5 (right). For the fixed schemes, this chart simply confirms that the aggregate transmission power of $19 \times p^{\star}_{\text{HS-DSCH}}$ Watt is (nearly) completely shared by an average of 200 data calls per constellation, where a negligible discrepancy exists beyond the fourth decimal, due to the non-zero probability that a NodeB is idle with respect to data calls and thus 'wastes' its HS-DSCH resources. The expected HS-DSCH transmission power per data call for the AP scheme appears to be slightly higher than for the AR scheme. Hence the AP scheme appears to utilise the resources unused by the prioritised speech calls to a somewhat greater extent, which is probably due to the fact that under the AR scheme, a very poorly located UE consumes a disproportionate amount of power from its serving NodeB, allowing lower HS-DSCH powers used at other NodeBs, in order to establish the same uniform expected throughput.

7.7.3. ON THE ADAPTIVE SCHEMES' RELATIVE DATA PERFORMANCE

In order to identify the precise constellation aspects that influence the relative performance of the power- and rate-fair schemes, a number of further investigations have been carried out. The number of speech and data calls in each constellation is now fixed at 800 and 190, respectively, where the chosen number of data calls enables us to enforce a spatially homogeneous distribution of $190/19 = 10$ data calls per NodeB (within each cell, the calls are still randomly located). Although in *all* 10000 constellations randomly generated to evaluate this traffic mix, the expected data throughput under the AP scheme exceeds that under the AR scheme, the following example illustrates that the AP scheme does not *generally* outperform the AR scheme.

Example 7.1 Consider a network with two NodeBs ($\mathbb{B} = \{1, 2\}$) each serving a single data call only ($\mathbb{M}_{\text{speech}} = \emptyset$, $\mathbb{M}_{\text{data}} = \{1, 2\}$), and let $b_1 = 1$, $b_2 = 2$. The path gains are given by $G_{11} = G_{12} = G_{22} = 10^{-11}$ and $G_{21} = 10^{-15}$. Under the AP scheme both data calls are served at $p_{\max}$ Watt, which leads to data throughputs $R_1(\text{AP}) \approx 1091.354$ kbits/s and $R_2(\text{AP}) \approx 3813.571$ kbits/s, so that $R^\star(\text{AP}) \approx 2452.462$ kbits/s. The bisection search procedure required to optimise the AR scheme yields $R_1(\text{AR}) = R_2(\text{AR}) = R^\star(\text{AR}) \approx 3556.700$ kbits/s for $p_1 = p_{\max}$ Watt and $p_2 \approx 0.460$ Watt. Observe that $R^\star(\text{AP}) < R^\star(\text{AR})$, which is due to the fact that under the AP scheme, data call 1 suffers greatly from NodeB 2's $p_{\max}$, while the results of the AR scheme show that data call 2 can still get a nearly as high throughput (as under the AP scheme) with a tiny fraction of the power, simply because it hardly experiences any interference from NodeB 1, but primarily from its own signal's orthogonality loss.

Example 7.2 Consider the constellation described in the previous example, but now with UE 1 on a less fortunate location, e.g. inside a building, with $G_{11} = G_{12} = 10^{-15}$. Under the AP scheme both data calls are served again at $p_{\max}$ Watt, which leads to data throughputs $R_1(\text{AP}) \approx 168.701$ kbits/s and $R_2(\text{AP}) \approx 3813.571$ kbits/s (as in Example 7.1), so that $R^\star(\text{AP}) \approx 1991.136$ kbits/s. The bisection search procedure required to optimise the AR scheme yields $R_1(\text{AR}) = R_2(\text{AR}) = R^\star(\text{AR}) \approx 189.620$ kbits/s for $p_1 = p_{\max}$ Watt (as in Example 7.1) and $p_2 \approx 0.002$ Watt. Observe that now $R^\star(\text{AP}) \gg R^\star(\text{AR})$, which is due to the fact that data call 2 can be assigned a high transmission power without causing any harm to data call 1, which in turn cannot achieve a high throughput even if there is no inter-cellular interference as its C/I is noise-dominated (verify that data call 1 gets a throughput of only 189.626 kbits/s even if NodeB 2 is silent). Clearly, the AP scheme exploits such a constellation much more efficiently than the AR scheme, which brings the favourably located data call down to the poor QOS level that is achievable by the badly located data call. Note that the worst path gain among the serving radio links is 10^{-15} (G_{11}), which is very small compared to the worst path gain of 10^{-11} (G_{11}) in Example 7.1.

As suggested by the above examples, one principal constellation aspect that influences the relative performance of the power- and rate-fair scheduling schemes is the presence of a relatively poor serving radio link. In fact, based on a targeted investigation of a number of specific scenarios outlined below, the path gain variability of the serving radio links of the data calls is identified as the *key* constellation characteristic that determines the relative performance of the AP and AR schemes.

Consider the following four scenarios which gradually incorporate the different sources of variability that underlie the heterogeneity of the constellations, i.e. *(i)* the presence of speech calls; *(ii)* the spatial heterogeneity of data calls; and *(iii)* the path gain variability of the data calls:

SCENARIO I:	no speech traffic	spatial homogeneity	path gain homogeneity
SCENARIO II:	with speech traffic	spatial homogeneity	path gain homogeneity
SCENARIO III:	with speech traffic	spatial homogeneity	path gain homogeneity
SCENARIO IV:	with speech traffic	spatial heterogeneity	path gain heterogeneity

where the spatial homo/heterogeneity and the path gain homo/heterogeneity refers to the data calls. In cases of data call path gain homogeneity, the path gain of a UE to a serving (interfering) NodeB is fixed to the *average* path gain on a serving (interfering) radio link as experienced in a ('normal') scenario with randomly sampled path gains. Note that SCENARIO IV corresponds with the realistic scenario as considered in the experiments discussed earlier in this section, but with $(M_{\text{speech}}, M_{\text{data}})$ fixed at $(800, 190)$.

For all four scenarios, Figure 7.6 plots the expected data throughputs $\mathbf{R}_{\text{data}}\,(\text{AP})$ and $\mathbf{R}_{\text{data}}\,(\text{AR})$ achieved by the AP and AR schemes, respectively, as well as the ratio $\mathbf{R}_{\text{data}}\,(\text{AP})\,/\mathbf{R}_{\text{data}}\,(\text{AR})$, for 10000 snapshots (each plotted bullet represents the result obtained for a single snapshot). The plots should be considered in a clockwise fashion, starting with SCENARIO I in the top left corner.

Since the snapshots as well as all data calls within each snapshot are indistinguishable in SCENARIO I (there are no random aspects included), the achieved data throughputs are all identical and there is trivially no performance difference between the AP and AR schemes. Adding the speech traffic in SCENARIO II not only significantly lowers the data QOS but also induces a clear discrepancy between the performance of the different scheduling schemes. Note that the expected throughput ratio is tightly bounded by unity, i.e. many constellations still allow similar throughputs under both schemes. The plot corresponding to SCENARIO III (allowing spatial data call heterogeneity) is comparable to that of SCENARIO II albeit that there is slightly greater variability in $\mathbf{R}_{\text{data}}$ (AP) $/\mathbf{R}_{\text{data}}$ (AR). As is clear from the final plot, corresponding to the 'complete' and realistic SCENARIO IV, the impact of the path gain heterogeneity on the relative performance of the adaptive schemes is most significant, raising the range of expected throughput ratios high above unity.

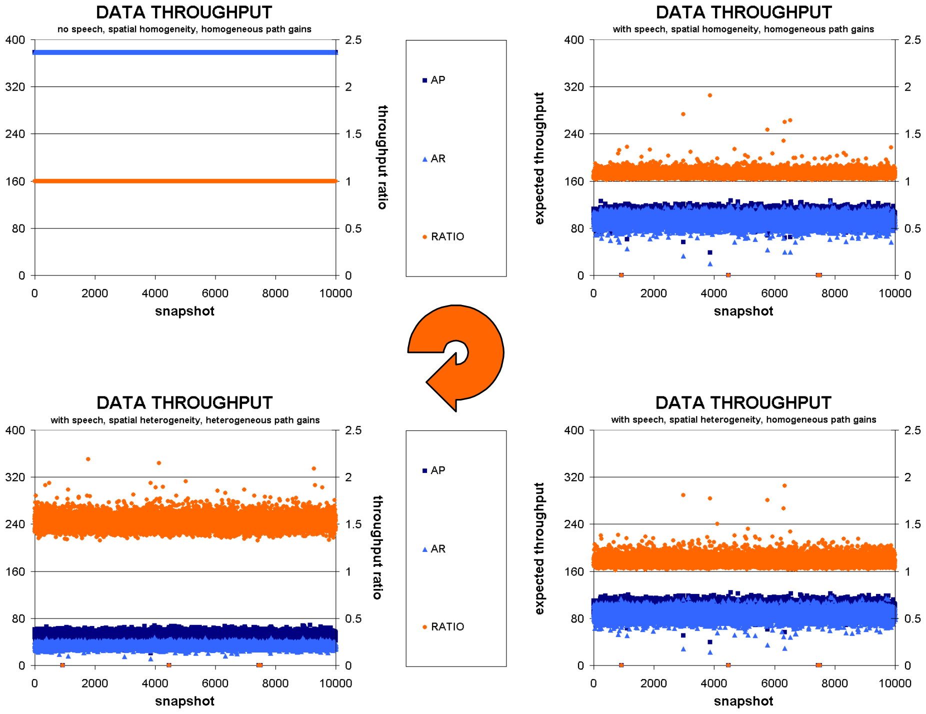


Figure 7.6 A comparison of the absolute and relative data performance for the AP and AR scheduling schemes. Four scenarios are considered, distinctly specified by the presence/absence of speech calls, the spatial homo/heterogeneity of data calls, and the homo/heterogeneity of the path gains on the data calls' serving radio links.

The observed large impact of the path gain variability on the data calls' radio links is demonstrated even more clearly by Figure 7.7 below. For each of 10000 snapshots, the figure depicts $\mathbf{R}_{\text{data}}\left(\text{AP}\right)$, $\mathbf{R}_{\text{data}}\left(\text{AR}\right)$ and $\mathbf{R}_{\text{data}}\left(\text{AP}\right)/\mathbf{R}_{\text{data}}\left(\text{AR}\right)$ as a function of the worst path gain among the serving data call radio links. The left chart is based on a single cell, data-only (20 calls) model with a fixed scheduling scheme (FP,FR), while the right chart corresponds with the 'normal' case of 19 cells with mixed speech (800 calls)/data (190 calls) traffic (cf. Figure 7.6) and compares the expected throughputs achieved by the adaptive scheduling schemes (AP,AR). Aside from the scatter plots both charts depict the conditional average of the performance measures within disjoint ranges in order to reveal the trend most clearly (white markers). As the left chart most clearly shows, the discrepancy between the expected data throughputs achieved by the scheduling schemes' different fairness objectives is largest if one or more data calls exist with an extremely poor serving radio link. The intuitive reasoning behind this lies in the excessive (minor) impact of a badly located data call on the overall data performance under the rate-fair (power-fair) scheme. Observe that if the worst path gain is rather large, implying that the radio link qualities are similar, the performance difference between the scheduling schemes becomes less significant. These trends are present but appear less significant in the multiple cell case, primarily due to the fact that the range of worst path gains is smaller given the large number of data calls.

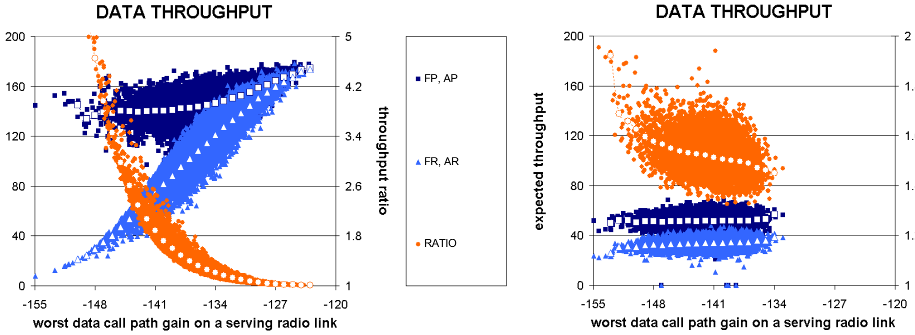


Figure 7.7 Absolute and relative data throughput performance under power- and rate-fair scheduling schemes versus the worst path gain on a serving radio link. The left chart depicts the results for a single-cell data-only scenario with fixed scheduling, while the right chart corresponds with a mixed services multi-cellular scenario with adaptive scheduling.

7.7.4. IMPACT OF THE TRAFFIC LOAD

Figure 7.8 shows the speech (left) and data (right) performance, respectively, of the different scheduling schemes for different traffic loads $(\rho_{\text{speech}}, \rho_{\text{data}}) = (\rho_{\text{speech}}, \frac{1}{4}\rho_{\text{speech}})$, where the value of ρ_{speech} is shown on the horizontal axis. The potential of adaptive schemes to improve the speech performance increases significantly as the growing load provides greater room for improvement, while at some point the speech traffic load becomes so high that virtually all speech calls experience outage, regardless of the presence of data calls, at which point adaptive scheduling can no longer 'save' the speech calls.

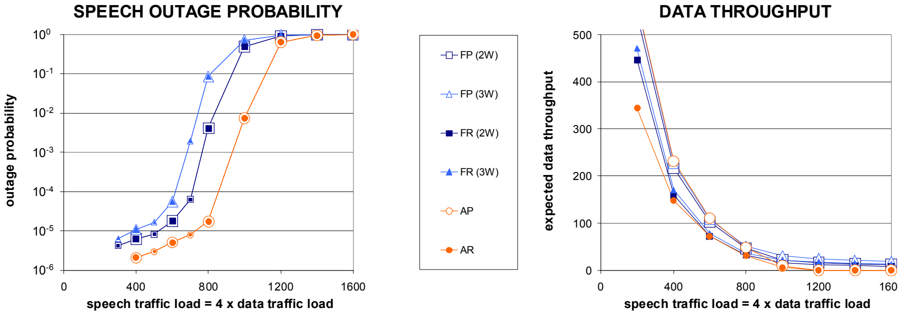


Figure 7.8 Speech call outage probabilities (left) and the expected data throughput (right) as a function of the traffic load for a given traffic mix of $\rho_{\text{speech}} : \rho_{\text{data}} = 4$.

The right chart indicates that the AP-curve crosses the FP-curves around $(\rho_{\text{speech}}, \rho_{\text{data}}) = (800, 200)$. Below this 'critical' traffic load, there is sufficient interference slack in the network to be exploited by the AP scheme's adaptivity and converted into higher data throughputs, while above the critical traffic load, the worsening speech performance demands the AP scheme to reduce the data QOS in support of the speech calls. Observe that whereas the adaptivity of the AP scheme provides a significant improvement of the speech performance due to enforced strict priority for the speech service, the improvement of the data QOS is never striking. Observe that the data QOS under the AR scheme lies below all other curves over the entire range of traffic loads. Apparently the AR scheme's task to protect speech calls as well as badly located data calls (an implicit consequence of the rate fairness objective) is so hard that it can only improve the speech QOS unilaterally (the primary task of adaptive schemes), and not the data QOS as well (the secondary task of adaptive schemes). Though it is beyond the numerical range of Figure 7.8 (right), for extremely low traffic loads, the expected data throughput under both adaptive schemes converges towards the upper bound given by

$$\frac{R_c}{\gamma_{\text{data}}} \frac{p_{\max}}{\omega p_{\max}} \approx \frac{3840}{10^{4/10}} \frac{15.849}{0.400 \cdot 15.849} \approx 3821.829 \text{ kbits/s},$$

a value that can be approached only by solitary data calls at prime locations.

7.7.5. QUALITATIVE COMPARISON OF SCHEDULING SCHEMES

In a high-level qualitative comparison of the considered schemes, we note that the principal advantage of the FP and FR schemes is the ease of their *implementation*. These fixed schemes only require the availability of fast link adaptation, while the FR scheme is slightly more involved due to the determination of the WRR packet scheduling weights. The AP and AR schemes are inherently more complex due to their adaptive nature, where in particular the close integration of rate control and packet scheduling and the required bisection search complicate the AR scheme significantly. With regard to the provided *resource efficiency*, the adaptive schemes are inherently most efficient. Although a conclusive comparison of the (non-)adaptive power- and rate-fair schemes requires further investigations (see also below), a reserved initial impression is that the power-fair schemes establish a slightly higher resource efficiency (in line with the conclusions in [158]), since they are less affected by badly located data calls. Finally, regarding the achieved rate and power fairness, the schemes operate in accordance with the posed objectives, noting that e.g. the FR scheme provides a lower degree of rate fairness than its adaptive equivalent (AR scheme). Table 7.3 summarises these qualitative comparisons.

While the presented results and the intuitive support seem to advocate the use of power- over rate-fair schemes, the implications of the spatial unfairness caused by power-fair scheduling are to be investigated further. The applied Monte Carlo evaluations are useful in demonstrating the potential gain of adaptive scheduling for power- and rate-fair scheduling schemes separately, with respect to their non-adaptive alternatives, and to provide an initial qualitative comparison of the different effects power- and rate-fair schemes have on the spatial throughput distribution. A conclusive comparison of both scheduling types must take into account the further evolution of

Table 7.3 A qualitative comparison of the investigated scheduling schemes with regard to their implementability, achieved resource efficiency and ability to establish rate (QOS) or power (resource) fairness.

	FP	FR	AP	AR
IMPLEMENTABILITY	++	+	-	- -
RESOURCE EFFICIENCY	-	- -	++	+
RATE FAIRNESS	- -	+	- -	++
POWER FAIRNESS	++	- -	++	- -

the considered data calls, as power-fair scheduling tends to dehomogenise the spatial distribution of data calls, favouring near over remote calls. As noted earlier, such a conclusive comparison requires an inevitably time-consuming integration of dynamic simulations and analytical optimisation.

7.8. CONCLUDING REMARKS

We have presented a performance evaluation of fair scheduling schemes in the CDMA downlink of an integrated services UMTS network serving prioritised speech calls on DCHs and data calls on HS-DSCHs involving Monte Carlo simulations in combination with analytical optimisation. A distinction was made between fixed and adaptive schemes on the one hand, and between fair distribution of the available power resources and fair QOS provisioning on the other hand. The principal objective was to demonstrate the potential performance enhancement that different adaptive scheduling schemes can establish for both service types, if the delay-tolerance and flexibility of the data service is exploited by up(down)grading the data rates in incidences of relatively light (heavy) speech traffic. Our results indicate that it is indeed possible to enhance the QOS of both speech and data services simultaneously by adaptive scheduling. Speech calls enjoyed the most significant performance gain due to the strict prioritisation policy. An initial comparison of power- and rate-fair schemes reveals that schemes of the former type strongly favour near over remote data calls, whereas schemes of the latter type provide spatially more homogeneous QOS levels. Although the overall data QOS seems to be higher under the power-fair schemes, the implications of the inherent spatial unfairness as the system evolves further in time, require further investigations in order to establish a conclusive comparison.

A number of other relevant research issues are recommended for future study. Based on the presented results, and given the significant performance enhancements that can be achieved by adaptive scheduling, the next step is to further assess the schemes' implementability, and potentially devise and evaluate heuristic derivatives of the investigated idealised schemes. For instance, the global scope of the adaptive scheduling schemes serves well as a reference scenario, while more practical schemes are likely to operate in a local fashion. Adaptive scheduling within localised NodeB clusters also relaxes the impact of a data traffic hot spot or a single poorly located data call on the overall data performance. The appropriate scope of an implemented scheme is a trade-off between performance and implementability (or speed of operation), and is co-determined by the network operator's policy with respect to fairness provisioning. Also, relief of the fairness requirement opens up possibilities for different data scheduling schemes, which may potentially achieve higher resource efficiencies at the cost of e.g. a greater data QOS variability. As the adaptivity gain of the considered scheduling schemes has been demonstrated to enhance primarily the speech QOS and the data QOS to a much lesser extent, the potential of a hybrid scheme, which assigns a minimum transmission power to non-idle HS-DSCHs and thus shares the adaptivity gain in a less extreme manner, should be assessed (see e.g. [150]). Further, a perhaps rather academic question remains whether the capacity slack left over by prioritised speech calls can be properly defined, and whether a unique adaptive scheduling scheme exists that utilises these resources most efficiently. In any case, in order to evaluate implementable (heuristic) adaptive scheduling schemes in a more realistic setting, dynamic simulations are advised, that include the mutual influence of scheduling and the dynamics of call arrivals and departures, the effects of the induced spatial traffic inhomogeneity, as well as the response of TCP flow control to the inherent throughput fluctuations. A first attempt of a dynamic simulation study involving adaptive scheduling is presented in [68].

CHAPTER 8

THE IMPACT OF MOBILITY ON UMTS NETWORK PLANNING

STOCHASTIC analysis of the intrinsic effects of terminal mobility on UMTS network performance seems to be a generally disregarded territory, in stark contrast with the multitude of similar studies that have been carried out for FDMA-based (first- or) second-generation networks. The principal complications in UMTS network performance studies are due to the underlying CDMA technology. Unlike in second-generation systems such as GSM, the universal frequency reuse in UMTS networks implies that the actual capacity of a given cell is directly linked to the loading of adjacent cells, which necessitates the appropriate consideration of a multi-cellular network model.

In a comparatively 'simple' multi-cellular scenario with a single circuit-switched service such as speech or video telephony, this chapter concentrates on the impact of terminal mobility on UMTS network planning, which stems from two distinct aspects that occur at different time scales. At the burst/packet level, a higher terminal velocity leads to more stringent energy-per-bit to interference-plus-noise density ratio requirements due to the combined effects of multipath propagation, Doppler shifts and Transmission Power Control imperfections, which in turn raises interference levels and thus reduces the network capacity. At the call level, a higher degree of terminal mobility necessitates a greater Radio Resource Reservation in support of call handovers, in order to prevent excessive call dropping. As a consequence, the inevitably raised fresh call blocking probability induces a need for denser site planning.

Following a decomposed evaluation and optimisation approach that is characterised by a segregation of the interference aspects and the traffic dynamics, the pursued impact of terminal mobility on network planning is assessed by deriving the minimum number of required NodeBs to cover a given area under some predetermined restrictions on outage, blocking and dropping. A conversion of the optimal number of NodeBs to implied network investment costs is applied in order to reveal the bottom-line impact of terminal mobility.

The outline of this chapter is as follows. Section 8.1 reviews some related literature, followed by a statement of the chapter's contribution in Section 8.2. In Section 8.3 the precise objective of the presented study is formulated in terms of investment costs and performance targets. Section 8.4 describes the applied model. A convenient yet for our purposes admissible abstraction of a realistic network is made in Section 8.5 regarding the formulation of system capacity. Section 8.6 then describes the mathematical analysis and optimisation procedure used to achieve the formulated objectives. Numerical results are presented and discussed in Section 8.7, while Section 8.8 ends the chapter with some concluding remarks.

8.1. LITERATURE

A large number of investigations exist regarding the impact of terminal mobility on network planning, generally concentrating on FD(/TD)MA-based cellular networks. The common focus in these studies is on the influence of mobility-induced handovers. An important distinction that can be made concerns the considered policy in giving preference to handover over fresh call requests [212]. The 'guard channel' concept comprises of a reservation of channels in each cell that are available to handover calls only, while the remaining channels are shared by both call types [44, 99, 108, 144, 165, 212, 229]. An either supplemental or substitutional policy is the queueing of handover requests, which exploits the common cellular overlap in the network. In case of a queued handover request, the considered call remains connected to its original base station until either the requested handover is accepted or its signal quality has deteriorated to the extent that the radio link is released . Under such policies, the experienced service quality in terms of the forced call termination probability can be further enhanced by devising an appropriate queueing discipline [44, 108, 212, 229]. Lastly, [179] suggests to allow the opportunity of 'directed retry', where a rejected call reattempts admission in an adjacent cell, only to handover calls.

Another distinction concerns the manner in which terminal mobility is included. In multi-cellular network analyses [36, 165, 179] terminals are randomly routed between different cells. The performance evaluation of the considered Markovian models either relies on fixed-point methods [165] or on product-form approximations [36, 179]. The multitude of analyses, however, approximate the network with a single cell model.

One common approach to model mobility in a single cell scenario is to assume an autonomous Poisson arrival process of handover requests [144, 229], allowing a straightforward analysis under Markovian assumptions [229] or suitable approximations in case of generally distributed cell residence times and call durations [144]. Alternatively, the handover call arrival rate can be implicitly derived as a function of the fresh call arrival rate and some assumption on the cell residence time, using fixed-point methods [44, 84, 108].

In [99] the impact of mobility on CDMA network performance is evaluated in terms of call blocking and dropping. The proposed modelling approach includes the aspect of soft handovers, where a call maintains a radio link with multiple serving NodeB's, albeit in a somewhat rudimentary manner. Unfortunately, the CDMA-specific dependencies between a given cell's actual loading and its adjacent cells' available capacity, as a consequence of co-channel interference, is neglected.

In [136] the impact of terminal mobility on the aggregate call request rate, comprising of newly originating calls and handover requests, is assessed for both traffic concentration and dispersion scenarios. A set of numerical experiments illustrate that any reasonable degree of terminal mobility imposes a significant increase of the aggregate call request rate, which should thus be incorporated in the network planning process.

8.2. CONTRIBUTION

The objective of the presented study is to provide insight regarding the impact of terminal mobility on UMTS radio network planning. To this end we develop and analyse a tractable two-stage model that is simple enough to allow true optimisation within reasonable time, while still sufficiently realistic to obtain valuable *qualitative* insights for network planning purposes, as it captures the UMTS network characteristics that are essential to our objectives, i.e. terminal mobility and inter-cellular dependencies. Analogous to some extent to the decomposed models applied in Chapters 6 and 9, the relevant interference aspects are incorporated in a conversion of the CDMA-inherent soft capacity to hard capacity (STAGE I). The hard capacity is expressed in terms of an admissible region that is applied to the subsequent call level performance optimisation of STAGE II. On the basis of this analysis, the investigation identifies terminal mobility as a key property to be taken into account in the planning process. It does so by deriving *bottom-line* performance measures, i.e. investment costs and service quality, which is precisely the relevant trade-off. The results further indicate the consequences of *not* properly incorporating terminal mobility in the planning process, again focussing on the bottom-line performance measures. The presented numerical trends are intended to assist network operators in developing planning guidelines. Additionally, the results provide a basis for more efficient and targeted numerical studies required for the actual planning process, indicating the relevant parameter regime and providing a means for verification.

8.3. OBJECTIVES

The objective of the presented investigation is to quantify the impact of terminal mobility on UMTS radio network planning. This impact stems from two distinct aspects.

- More severe E_b/N_o *requirements* apply in case of higher terminal velocities, in order to achieve a given target BLock Error Rate (BLER), due to the combined effect of multipath propagation, Doppler shifts and Transmission Power Control imperfections. In principle fast fading affects the BLER of faster terminals to a lesser extent, as deep fades are brief and thus so are erroneous bit clusters, which are therefore more easily corrected thanks to the implemented interleaving scheme. However, link level simulations indicate that the opposing effects of Doppler shifts and Transmission Power Control imperfections outweigh the interleaving gain (e.g. [78]).
- The greater the degree of mobility, the greater the *radio resource reservation* regarding handovers should be, in order to keep the dropping probability $\mathbf{P}_d$ below a prespecified target value. Furthermore, the larger this reservation, the smaller the capacity that remains to serve fresh calls and thus the higher the blocking probability $\mathbf{P}_f$. As a consequence, denser site planning (i.e. a smaller radius R) is required to reduce the fresh call blocking probability $\mathbf{P}_f$ in order to comply with the prespecified target value. Note that as a result of denser site planning, i.e. smaller cells, the degree of mobility and thus the handover rates increase (see Figure 8.1).

The objective of a network operator is to minimise the investment costs, i.e. to maximise the cell radius R, such that both its Quality Of Service target, defined as the maximum allowed outage probability, denoted $\mathbf{P}_o^\star$, and its Grade Of Service targets,

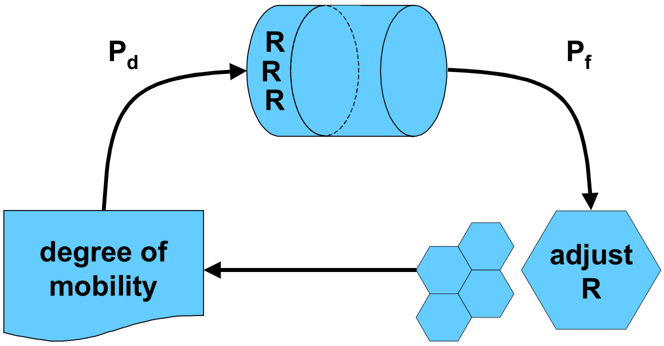


Figure 8.1 Schematic overview of the impact of terminal mobility on UMTS network planning due to the need to reserve resources in anticipation of handovers.

denoted $\mathbf{P}_f^\star$ and $\mathbf{P}_d^\star$, are met. Stated formally, the objective is as follows:

$$\text{OBJECTIVE:} \quad \max\ R \ \text{ subject to } \mathbf{P}_o \leq \mathbf{P}_o^\star \text{ and } \mathbf{P}_f \leq \mathbf{P}_f^\star \text{ and } \mathbf{P}_d \leq \mathbf{P}_d^\star,$$

where besides R, the Radio Resource Reservation level RRR and the CAC-related state space $\mathbb{S}$ of feasible network states (see below) are the optimisation parameters. The relevant performance measures will be defined in more detail below.

In correspondence with the layered model of capacity allocation presented in Chapter 1, note that in the current model CAC is deployed to satisfy the QOS target ($\mathbf{P}_o$) by an appropriate choice of $\mathbb{S}$, RRR is used to balance $\mathbf{P}_f$ and $\mathbf{P}_d$, and proficient network planning (choice of R) ensures that the absolute GOS levels meet their individual target values.

8.4. MODEL

This section describes the system, propagation and traffic aspects of the model that is used for the investigation of the impact of terminal mobility on UMTS traffic management and network planning. In addition, the relevant performance measures are specified.

8.4.1. SYSTEM MODEL

The system model that is proposed for the evaluation of the impact of terminal mobility on UMTS network performance is an $M \times M$ cellular network, illustrated by Figure 8.2 (left) for $M = 2$. In order to approximate an infinite network a reference cellular area of $B \equiv M \times M$ hexagonal cells is wrapped around into a torus-shape. In this torus-shaped network the (numbered) reference cells are repeated in a regular manner, as indicated. As a result, each cell has neighbours on all sides so that mobile terminals remain within the network model as they cross an imaginary border at the edge of the reference area and interference is experienced from all directions. Denote with $\mathbb{B} \equiv \{1, \cdots, B\}$ the set of NodeBs (or cells) and let $\mathbb{B}_b$ denote the set of NodeBs adjacent to NodeB $b \in \mathbb{B}$.

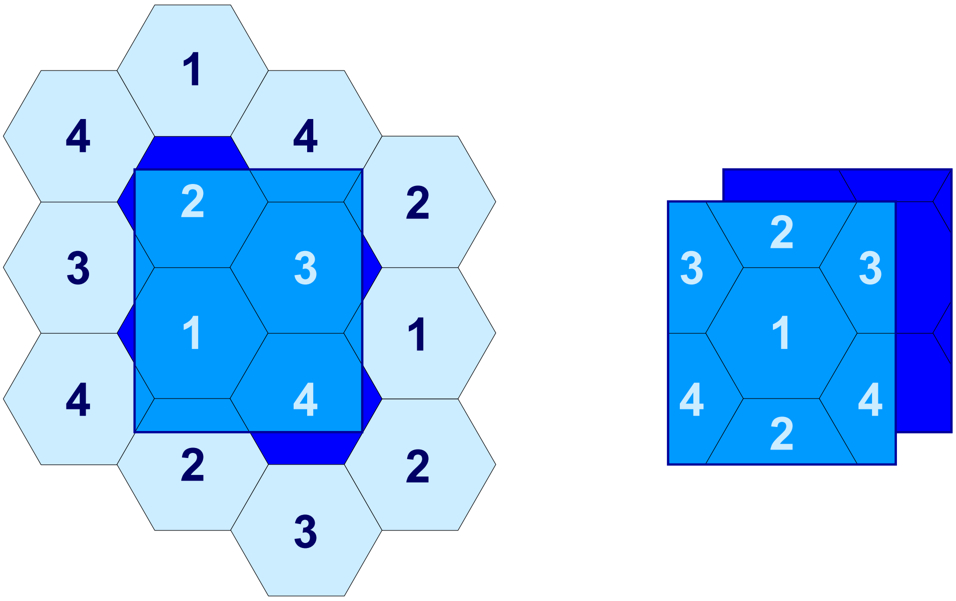


Figure 8.2 Illustration of the 2 × 2 cellular network model. Wraparound techniques are used to mimic an infinite network in order to avoid boundary effects with regard to terminal mobility and signal interference.

Since a given terminal is not only located in one of the reference cells, but 'shadows' of this terminal also recur in all lightly shaded copies of this cell, there is a degree of freedom in choosing which of all these 'versions' of the considered terminal is to be considered in the interference calculation for a given NodeB. As a worst-case approach (to partially correct the absence of additional cell tiers for small M) we select either the original terminal or one of its shadows, whichever version has the lowest path loss to the considered NodeB, i.e. the nearest one in a case without shadow fading. As a consequence, all interference experienced by e.g. NodeB 1 originates from the rectangular region indicated in Figure 8.2 (right). Observe that the shape and orientation of the semi-hexagons of the adjacent cells differ.

As the presented investigation concentrates on (symmetric) telephony services (see below), in contrast to the data performance evaluations of Chapters 6 and 7, the impact analysis concentrates on the *uplink* direction of transfer, which is the logical bottleneck direction given the lack of signal orthogonality that is due to the inherent asynchrony of the uplink transmitters, as well as the generally more limited terminal transmission power budget.

8.4.2. PROPAGATION MODEL

Ideal uplink Transmission Power Control is assumed with equal received power levels in each NodeB, while a simple distance-based propagation model is assumed to relate the transmission power $p_{\text{transmission}}$ and the received power $p_{\text{reception}}$, given by

$$p_{\text{reception}} = p_{\text{transmission}} \cdot r^{-\varsigma},$$

with attenuation exponent ς, and r the distance between transmitter and receiver. No thermal noise is included, so that the received power level $p_{\text{reception}}$ can be fixed at 1 without loss of generality.

It is noted that in the considered scenario without slow or fast fading fluctuations, the potential gains from macro-diversity as attainable in an actual CDMA network, are captured by the assumption that each terminal is served by the nearest (best) NodeB, given the generally applied ‘selection combining’ method in the CDMA uplink, where a serving RNC chooses the best of potentially multiple signals that may be detected by its associated NodeBs.

8.4.3. TRAFFIC MODEL

The considered UMTS network serves calls of a single type (either speech or video telephony), that are generated according to a spatially uniform Poisson arrival process at a given nominal rate of λ_o calls/$\big($second $\cdot$ km$^2\big)$, and have exponentially distributed call durations with mean μ^{-1}. The nominal traffic load, expressed in Erlang/km^2, is equal to $\rho_o \equiv \lambda_o/\mu$. All terminals are assumed to have a uniform velocity v which is either 3, 50 or 120 km/h, and move in straight but random direction. The service- and velocity-specific energy-per-bit to interference-plus-noise density ratio (E_b/N_o) target is denoted $\gamma_v^\star$, while the corresponding carrier-to-interference ratio (C/I) target is denoted $\widetilde{\gamma}_v^\star \equiv \gamma_v^\star R_i/R_c$, with R_c the system chip rate and R_i the information bit rate associated with the considered service. The E_b/N_o target values correspond with the included BLER target values. The call activity factor is denoted by α. Table 8.1 summarises the relevant service parameters (see also [219]). Some of the parameters are based on the evaluation report of the α-proposal [78], as it proved a valuable source for the assumed service and velocity-dependent E_b/N_o target values.

Table 8.1 Overview of the numerical settings of the relevant service characteristics.

	SPEECH		VIDEO	
μ^{-1}	100	seconds	100	seconds
α	0.4		1.0	
R_c	3840	kchips/s	3840	kchips/s
R_i	12.2	kbits/s	384	kbits/s
BLER*	10^{-3}		10^{-6}	
$\gamma_3^\star$	3.3	dB	2.2	dB
$\gamma_{50}^\star$	4.0	dB	2.6	dB
$\gamma_{120}^\star$	5.0	dB	3.1	dB

8.4.4. PERFORMANCE MEASURES

The performance of the speech and video services at the physical layer is expressed in terms of the *outage probability* $\mathbf{P}_o$, i.e. the probability that an arbitrary call fails to meet its C/I requirement. At the call level, the *fresh call blocking probability* $\mathbf{P}_f$ and *call dropping probability* $\mathbf{P}_d$ are considered as the principal GOS measures.

The former measure specifies the probability that an arriving call is denied access to the network, while the latter measure gives the probability that an admitted call terminates prematurely, as a consequence of a failed handover attempt.

8.5. SYSTEM CAPACITY

A CDMA network is characterised by *soft capacity* in the sense that there is no hard limit on the number of concurrent calls that can be served. While it is in principle always possible to admit an additional call to a CDMA network, this flexibility comes at a cost of a graceful degradation of the radio link qualities (QOS, BLER) of existing calls. In practice, a network operator will pose a minimum requirement on the link quality of active calls, which induces a need for Call Admission Control and thus converts soft into *hard capacity*. This conversion constitutes STAGE I of our analysis.

Consider network state (n_r, n_a) as seen from a reference cell where n_r and n_a denote the (aggregate) number of calls in the reference and adjacent cells, respectively (only intra- and adjacent-cell interference is taken into account). The QOS experienced by the calls in the reference cell is defined by the *outage probability* $\mathbf{P}_o\,(n_r, n_a)$, i.e. the probability that an active call does not meet its E_b/N_o target. In order to derive an expression for $\mathbf{P}_o(n_r, n_a)$, the amount of interference experienced in the reference cell from the n_a calls in the adjacent cells is well-approximated by a Gaussian distributed random variable $I_a\,(n_a)$ taking into account randomness caused by terminal location and call activity (see e.g. [79, 91]).

Using standard techniques we now determine the mean and variance of $I_a(n_a)$. The approach is analogous to that followed in e.g. [91], albeit slightly modified in accordance with the considered torus-shaped network model. Figure 8.3 shows a UE with polar coordinates (r, θ) with respect to its serving NodeB, in a cell adjacent to the reference cell at an inter-NodeB distance $D = R\sqrt{3}$, with R the cell radius. Using straightforward trigonometry, the distance from the remote UE to the reference NodeB is given by $\sqrt{D^2 - 2\,D\,r\,\cos\theta + r^2}$, so that the amount of interference $p_{\text{interference}}$ received from this UE at the reference NodeB equals

$$p_{\text{interference}} = p_{\text{transmission}} \left(\sqrt{D^2 - 2\,D\,r\,\cos\theta + r^2}\right)^{-\varsigma},$$

where $p_{\text{transmission}} = r^{\varsigma}$ is the mobile's transmission power assuming perfect Transmission Power Control and a received power target of 1.

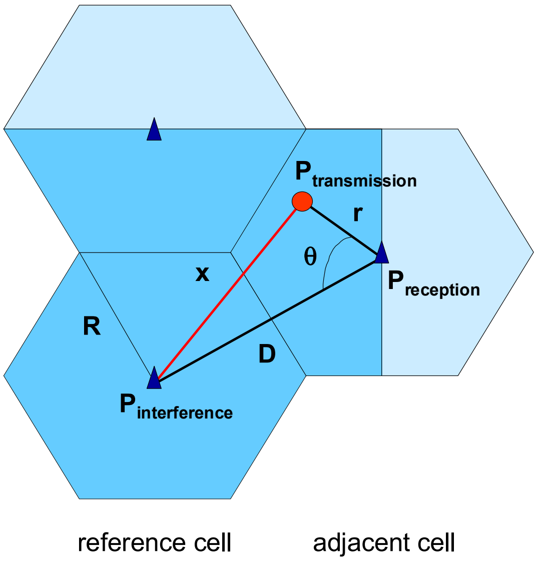


Figure 8.3 Calculation of the expectation and variance of the adjacent cell interference in the considered torus-shaped UMTS network.

Given a uniform spatial distribution (in Cartesian coordinates) of UEs over the considered cell area, and approximating the adjacent semi-hexagons by semi-circles, the k^{th} moment of the random amount of interference received from a single UE in an adjacent cell is given by

$$\mathbf{E}_{\theta^\star}\left\{p^k_{\text{interference}}\right\} = \int\limits_{r=0}^{R}\int\limits_{\theta=\theta^\star}^{\theta^\star+\pi}\left[r^{\varsigma}\left(\sqrt{D^2 - 2\,D\,r\,\cos\theta + r^2}\right)^{-\varsigma}\right]^k \cdot \frac{r}{\pi R^2}\,d\theta\,dr,$$

with $\theta^\star = -\frac{\pi}{3}$ for the adjacent cell depicted in Figure 8.3, and $\theta^* = -\frac{\pi}{2}$ for the slightly differently oriented adjacent cell just above the reference cell (see also Section 8.4.1). The above integral turns out to be independent of R (and hence of $D = R\sqrt{3}$: scalability). The expected value and variance of $p_{\text{interference}}$ are then appropriately determined as follows:

$$\mathbf{E}\left\{p_{\text{interference}}\right\} = \alpha\left(\frac{2}{3}\mathbf{E}_{-\frac{\pi}{3}}\left\{p_{\text{interference}}\right\} + \frac{1}{3}\mathbf{E}_{-\frac{\pi}{2}}\left\{p_{\text{interference}}\right\}\right),$$

and

$$\begin{aligned}\mathbf{V}\left\{p_{\text{interference}}\right\} &= \alpha\mathbf{E}\left\{p_{\text{interference}}^2\right\} - \left(\alpha\mathbf{E}\left\{p_{\text{interference}}\right\}\right)^2 \\ &= \alpha\left(\frac{2}{3}\mathbf{E}_{-\frac{\pi}{3}}\left\{p_{\text{interference}}^2\right\} + \frac{1}{3}\mathbf{E}_{-\frac{\pi}{2}}\left\{p_{\text{interference}}^2\right\}\right) + \\ &\quad - \left(\alpha\mathbf{E}\left\{p_{\text{interference}}\right\}\right)^2,\end{aligned}$$

with α the call activity factor. Using $\mathbf{E}_{-\frac{\pi}{3}}\left\{p_{\text{interference}}\right\} = 0.2248$, $\mathbf{E}_{-\frac{\pi}{2}}\left\{p_{\text{interference}}\right\} = 0.2293$, $\mathbf{E}_{-\frac{\pi}{3}}\left\{p_{\text{interference}}^2\right\} = 0.2209$, $\mathbf{E}_{-\frac{\pi}{2}}\left\{p_{\text{interference}}^2\right\} = 0.2215$, the expected interference power $\mathbf{E}\left\{p_{\text{interference}}\right\}$ and its variance $\mathbf{V}\left\{p_{\text{interference}}\right\}$ are given in Table 8.2 below for $\alpha \in \{0.4, 1.0\}$ and $\varsigma = 3.64$.

Table 8.2 The expectation and variance of the adjacent cell interference.

	$\alpha = 0.4$	$\alpha = 1.0$
$\mathbf{E}\left\{p_{\text{interference}}\right\}$	0.0905	0.2263
$\mathbf{V}\left\{p_{\text{interference}}\right\}$	0.0803	0.1699

Observe that these values hold for a *single* randomly located terminal in an adjacent cell, while the mean and variance of $I_a(n_a)$ are readily derived, so that

$$I_a\left(n_a\right) \sim \begin{cases} N\left(0.0905\,n_a, 0.0803\,n_a\right) & \text{for } \alpha = 0.4, \\ \\ N\left(0.2263\,n_a, 0.1699\,n_a\right) & \text{for } \alpha = 1.0, \end{cases} \tag{8.1}$$

using $\varsigma = 3.64$ and the considered torus-shaped network model with $M = 2$. Given $n_r - 1$ other calls in the reference cell, the amount of intra-cell interference $I_r\left(n_r - 1\right)$ experienced by a given call in the reference cell is equal to the binomially distributed number of interfering calls in the reference cell,

$$I_r\left(n_r - 1\right) \sim \text{Binomial}\left(n_r - 1, \alpha\right),$$

as a result of the random call activity. The outage probability is then determined by conditioning on the number of active calls in the reference cell,

$$\begin{aligned}\mathbf{P}_o\left(n_r, n_a\right) &= \Pr\left\{\frac{1}{I_r(n_r-1)+I_a(n_a)} < \widetilde{\gamma}_v^{\star}\right\} \\ &= \sum_{n_r'=0}^{n_r-1} \binom{n_r-1}{n_r'} \alpha^{n_r'} \left(1-\alpha\right)^{n_r-n_r'} \overline{\Phi}_{\mu,\sigma^2}\left(\left(\widetilde{\gamma}_v^{\star}\right)^{-1} - n_r'\right),\end{aligned}$$

where $\overline{\Phi}_{\mu,\sigma^2}\left(\cdot\right)$ denotes the complementary Gaussian CDF, with mean μ and variance σ^2 taken from (8.1). For speech and video calls with velocity $v = 120$ km/h, Figure 8.4 (left) shows $\mathbf{P}_o\left(n_r, n_a\right)$ as a function of n_r and n_a, while Figure 8.4 (right) shows the corresponding iso-outage curves at levels 1%, 10%, 50%, 90% and 99%.

Posing an operator-specified maximum on the allowed outage probability of 1% enables us to convert the CDMA-specific *soft* capacity into a *hard* capacity, as the corresponding iso-outage curve specifies a set of feasible pairs (n_r, n_a), which is well-approximated with a linear inequality of the form $\kappa_r n_r + \kappa_a n_a \leq 1$, where κ_r (κ_a) denotes the *effective interference* of a call in the reference (adjacent) cell, respectively [79]. Applying this inequality to each cell as a reference cell, yields an B-dimensional state space of feasible system states

$$\mathbb{S} \equiv \{\mathbf{n} \equiv (n_1, \cdots, n_B) \in \mathbb{N}_0^B : n_b\kappa_r + \sum_{b' \in \mathbb{B}_b} n_{b'}\kappa_a \leq 1,\ b \in \mathbb{B}\},$$

where n_b denotes the number of calls served by NodeB $b \in \mathbb{B}$.

Table 8.3 contains the effective interference values for both services and all three terminal velocities ($M = 2$). Note that the effective interference of a video call is significantly larger than that of a speech call, which reflects the fact that video calls require higher bit rates and have a higher call activity factor. As a result, fewer video calls can be admitted to the network. Furthermore, the effective interference is increasing in the terminal velocity, as a result of the higher E_b/N_o requirement.

We note that this approach of deriving effective interference (or: bandwidth) values is common in fixed network performance analyses in deriving network capacities and Call Admission Control schemes, and has recently been introduced to CDMA networks (see [79]).

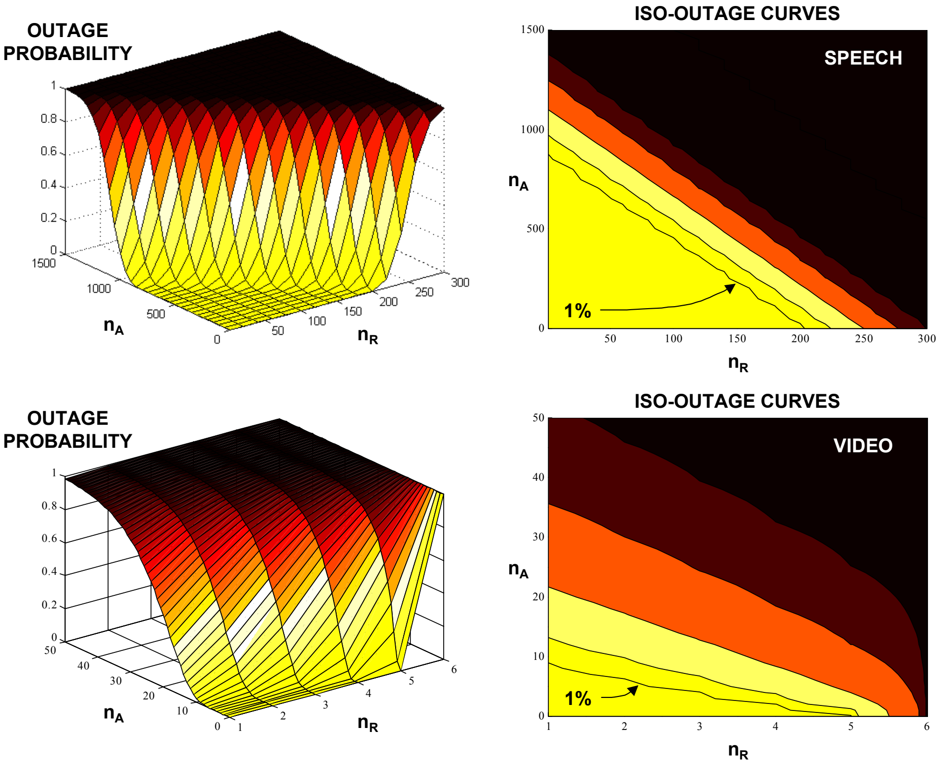


Figure 8.4 Speech (top) and video (bottom) call outage probabilities $\mathbf{P}_o(n_r, n_a)$ as a function of the number of calls n_r (n_a) in the reference (adjacent) cell(s) for a velocity of $v = 120$ km/h. The charts on the right depict the corresponding iso-outage curves (the slight irregularity of the iso-outage curves is due to intrapolations done by MATLAB).

8.5.1. CAC AND RRR

The derived state space $\mathbb{S}$ can be applied directly to execute Call Admission Control: if a fresh call or handover request originating in system state $\mathbf{n} \in \mathbb{S}$ is to bring the system in a new state $\mathbf{n}' \in \mathbb{S}$ the request is granted, otherwise the call is blocked or dropped, respectively. However, as call dropping is generally regarded worse than call blocking, *radio resource reservation* can be applied to protect existing calls, by means of reserving a subset of states in $\mathbb{S}$ that can only be reached as a result of call handovers. The proposed reservation scheme is implemented such that a fresh call requesting service from NodeB b in system state $\mathbf{n} \in \mathbb{S}$ is admitted if and only if

Table 8.3 The effective interference values for both services, the reference and adjacent cells, and three different terminal velocities.

	SPEECH		VIDEO	
v	κ_r	κ_a	κ_r	κ_a
3 km/h	0.0031497797	0.0007342144	0.1666666666	0.0724637681
50 km/h	0.0037445319	0.0008748906	0.1803278689	0.0819672131
120 km/h	0.0048027057	0.0011273957	0.2000000000	0.1000000000

$\mathbf{n} + (1 + RRR)\,\mathbf{e}_b \in \mathbb{S}$, where RRR denotes the Radio Resource Reservation level and $\mathbf{e}_b$ is a vector with a 1 on the b^{th} position and zeros elsewhere. Define

$$\mathbb{S}_b^\star \equiv \{\mathbf{n} \in \mathbb{S} : \mathbf{n} + (1 + RRR)\mathbf{e}_b \in \mathbb{S}\},$$

as the set of states where NodeB b admits fresh calls, while the complementary set $\mathbb{S}\backslash\mathbb{S}_b^\star$ contains all states where NodeB b blocks fresh call arrivals.

8.6. PERFORMANCE OPTIMISATION

This section describes the STAGE II call level performance optimisation procedure. A Markov chain is defined that models the system evolution. Subsequently, the relevant performance measures are derived and an outline is given of the implemented procedure to optimise network performance over the available control parameters, i.e. the cell radius R and the radio resource reservation level RRR.

8.6.1. MARKOV CHAIN

For given environment parameters λ_o, μ, v, and control parameters R and RRR, the system evolution can be described by an irreducible B-dimensional continuous-time Markov chain $(\mathbf{N}(t) \equiv (N_1(t), \cdots, N_B(t)))_{t\geq 0}$, with states denoted $\mathbf{n}$ and state space $\mathbb{S}$. The Markov chain's transitions are specified by fresh call arrivals, successful call terminations and call handover attempts:

- given cell radius R (in meters) and the nominal fresh call arrival rate of λ_o calls/second/m^2, the *fresh call arrival rate* per (hexagonal) cell is equal to $\lambda(\lambda_o, R) = \frac{3}{2}\lambda_o R^2\sqrt{3}$; the offered traffic load is given by $\rho(\lambda_o, R) \equiv \lambda(\lambda_o, R)/\mu$;

- call durations are assumed to be exponentially distributed with mean μ^{-1}, so that the *call termination rate* is equal to μ;
- cell residence times are assumed to be exponentially distributed with mean $\left(\frac{8}{3\pi}R\right) / \left(\frac{v}{3.6}\right)$ where $\frac{8}{3\pi}R$ is the average traversed distance of a randomly located user to the edge of its cell (see the intermezzo 'AVERAGE DISTANCE TO THE EDGE OF A CELL' below), so that the *call-specific handover request rate*, denoted $\zeta(v, R)$, is equal to the reciprocal of the mean, to be evenly split over the three possible target cells, i.e. $\zeta(v, R) = (\pi\, v) \,/\, (28.8\, R)$.

For $\mathbf{n}, \mathbf{n}' \in \mathbb{S}$, the Markov chain's transition rates are as follows:

$$\mathcal{Q}(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda(\lambda_o, R) & \text{if } \exists b \in \mathbb{B},\ \mathbf{n} \in \mathbb{S}_b^\star, \\ & \quad \mathbf{n}' = \mathbf{n} + \mathbf{e}_b, \\ n_b\mu + n_b\zeta\,(v, R)\,|\{b' \in \mathbb{B}_b : \mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'} \notin \mathbb{S}\}| & \text{if } \exists b \in \mathbb{B}, \\ & \quad \mathbf{n}' = \mathbf{n} - \mathbf{e}_b, \\ n_b\zeta(v, R) & \text{if } \exists b \in \mathbb{B},\ b' \in \mathbb{B}_b, \\ & \quad \mathbf{n}' = \mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'}, \end{cases}$$

where the different events correspond with an admitted fresh call, a terminating call (successful or dropped) and an admitted handover, respectively. All other non-diagonal entries of the infinitesimal generator $\mathcal{Q}$ are 0, while the diagonal entries are such that all rows of $\mathcal{Q}$ sum up to 0.

As the finite state space Markov chain is irreducible, a unique probability vector $\boldsymbol{\pi}$ exists that satisfies the system of global balance equations $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0}$ [213]. Efficient techniques exist to numerically solve this system and determine the equilibrium distribution $\boldsymbol{\pi}$, e.g. the applied successive overrelaxation method [213]. As the computational burden involved can still be rather significant, in particular for large state spaces, two model adjustments that lead to a product-form equilibrium distribution are outlined below, which may be used to swiftly determine an approximate solution.

INTERMEZZO: AVERAGE DISTANCE TO THE EDGE OF A CELL

We now determine the average distance of a randomly located terminal to the edge of its cell, which was applied in the above specification of the handover request rates.

Recall that terminals originate according to a spatially uniform Poisson process and are assumed to move in a straight line at a random angle. Consider a *circular* cell with radius R (see Figure 8.5). Let (r, θ) denote the polar coordinates of a randomly located terminal. It is readily verified that, without loss of generality, we can confine ourselves to assuming that the terminal is located in the upper right cell quarter and is moving in a straight horizontal line towards either the left or right, covering distances denoted Δ_1 and Δ_2, respectively. Straightforward trigonometry gives

$$\Delta_1 = r\cos\theta + \sqrt{R^2 - (r\sin\theta)^2} \quad \text{and} \quad \Delta_2 = -r\cos\theta + \sqrt{R^2 - (r\sin\theta)^2}.$$

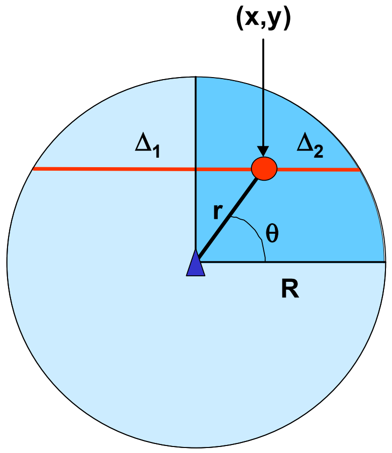


Figure 8.5 Calculation of the average distance of a randomly located UE to the edge of its cell, given a randomly selected direction of movement.

Under the assumption of uniformly distributed terminals (in Cartesian coordinates) over the considered cell quarter, the terminal location (r, θ) (in polar coordinates) can be derived to have probability density $4r/\left(\pi R^2\right)$, so that the average distance of a randomly located terminal to the edge of its cell is given by

$$\mathbf{E}\{\Delta\} = \int_{\theta=0}^{\frac{1}{2}\pi}\int_{r=0}^{R} \frac{4r}{\pi R^2}\left(\frac{\Delta_1 + \Delta_2}{2}\right) dr\, d\theta = \int_{\theta=0}^{\frac{1}{2}\pi}\int_{r=0}^{R} \frac{4r}{\pi R^2}\left(\sqrt{R^2 - (r\sin\theta)^2}\right) dr\, d\theta$$

$$
\begin{aligned}
&= \int_{\theta=0}^{\frac{1}{2}\pi}\int_{u=0}^{1} \frac{4}{\pi} uR \left(\sqrt{1-(u\sin\theta)^2}\right) du\, d\theta = \frac{4}{\pi}R\int_{\theta=0}^{\frac{1}{2}\pi} -\frac{1}{3}\frac{\left(1-\sin^2\theta\right)^{\frac{3}{2}}-1}{\sin^2\theta}\, d\theta \\
&= \frac{8}{3\pi}R,
\end{aligned}
$$

where the third equality is obtained after substitution of $u \equiv r/R$. Note that $\mathbf{E}\{\Delta\}$ is smaller than the cell radius R.

APPROXIMATIONS

We now discuss two distinct model adjustments proposed in [36] and [179] that lead to a product-form equilibrium distribution and are thus potentially functional in case of a large state space. As neither reference provides an indication of the accuracy of the proposed approximation, one of the numerical experiments presented in Section 8.7 will exploit the product-form approximation to demonstrate its merit.

Pallant and Taylor [179] propose to modify the original Markov chain by assuming that if a call's handover attempt fails, the call continues in its original cell. For $\mathbf{n}, \mathbf{n}' \in \mathbb{S}$, the modified Markov chain's transition rates are as follows:

$$
\mathcal{Q}_{\text{PT}}(\mathbf{n}, \mathbf{n}') \equiv \begin{cases} \lambda(\lambda_o, R) & \text{if } \exists b \in \mathbb{B},\ \mathbf{n} \in \mathbb{S}_b^{\star},\ \mathbf{n}' = \mathbf{n} + \mathbf{e}_b \\ \\ n_b\mu & \text{if } \exists b \in \mathbb{B},\ \mathbf{n}' = \mathbf{n} - \mathbf{e}_b, \\ \\ n_b\zeta(v, R) & \text{if } \exists b \in \mathbb{B},\ b' \in \mathbb{B}_b,\ \mathbf{n}' = \mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'}, \end{cases}
$$

where the different events correspond to an admitted fresh call, a successfully terminating call and an admitted handover, respectively. All other non-diagonal entries of $\mathcal{Q}_{\text{PT}}$ are 0, while the diagonal entries are such that all rows of $\mathcal{Q}_{\text{PT}}$ sum up to 0. Observe that the model modification eliminates handover blocking. For the case *without* Radio Resource Reservation, i.e. $RRR = 0$, the *product-form* equilibrium distribution of the reversible modified Markov chain is given by [179, Theorem 1]

$$
\widehat{\pi}_{\text{PT}}(\mathbf{n}) = \left(\sum_{\mathbf{n}\in\mathbb{S}}\prod_{b\in\mathbb{B}} \frac{\rho(\lambda_o, R)^{n_b}}{n_b!}\right)^{-1} \prod_{b\in\mathbb{B}} \frac{\rho(\lambda_o, R)^{n_b}}{n_b!},\ \mathbf{n} \in \mathbb{S}. \tag{8.2}
$$

Observe that the effect of terminal mobility in the form of $\zeta(v, R)$ does not appear in the product-form equilibrium distribution. This is due to the fact that in the modified Markov chain no calls are lost due to failed handovers: calls merely traverse the network until they successfully terminate. As also noted in [179, Theorem 1], for the case *with* RRR the modified Markov chain does *not* have a product-form equilibrium distribution. In order to still offer an approximation for the case that the reservation level $RRR > 0$, the equilibrium distribution in expression (8.2) can be applied, but then rescaled to the reduced state space $\widehat{\mathbb{S}}(RRR) \subseteq \mathbb{S}$ of *recurrent* states, in order to take into account that the deployment of an RRR scheme may make some states in $\mathbb{S}$ unachievable. Note that in the original Markov chain no such state space reduction needed to be enforced, as $\pi(\mathbf{n}) = 0$ for $\mathbf{n} \in \mathbb{S} \backslash \widehat{\mathbb{S}}(RRR)$ 'naturally' followed from solving the global balance equations). It is noted that the impact of the reservation level is further incorporated in the derivation of the performance measures below.

Boucherie and Mandjes [36] propose another modification of the original Markov chain: if a call's handover attempt fails, the call re-attempts in the target cell's adjacent cells at a redial rate r. For $\mathbf{n}, \mathbf{n}' \in \mathbb{S}$, the modified Markov chain's transition rates are then as follows:

$$\mathcal{Q}_{\text{BM}}(\mathbf{n}, \mathbf{n}') \equiv \begin{cases} \lambda(\lambda_o, R) 1\{\mathbf{n} \in \mathbb{S}_b^\star\} & \text{if } \exists b \in \mathbb{B}, \\ \quad + r\left|\{b' \in \mathbb{B}_b : \mathbf{n} + \mathbf{e}_{b'} \notin \mathbb{S}\}\right| & \quad \mathbf{n}' = \mathbf{n} + \mathbf{e}_b, \\ & \\ n_b \mu & \text{if } \exists b \in \mathbb{B}, \\ \quad + n_b \zeta(v, R)\left|\{b' \in \mathbb{B}_b : \mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'} \notin \mathbb{S}\}\right| & \quad \mathbf{n}' = \mathbf{n} - \mathbf{e}_b, \\ & \\ \qquad\qquad n_b \zeta(v, R) & \text{if } \exists b \in \mathbb{B},\ b' \in \mathbb{B}_b, \\ & \quad \mathbf{n}' = \mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'}, \end{cases}$$

where the different events correspond with an admitted fresh call or an admitted redialling call, a terminating call (successful or dropped) and an admitted handover, respectively. If $RRR = 0$, the redial rate r can be determined such that the equilibrium distribution is of product-form [36], which appears to be identical to that obtained via the modification proposed in [179, Theorem 1], i.e. $\widehat{\pi}_{\text{BM}}(\mathbf{n}) = \widehat{\pi}_{\text{PT}}(\mathbf{n})$, for $\mathbf{n} \in \mathbb{S}$. The treatment of the case where $RRR > 0$ is as outlined above.

8.6.2. PERFORMANCE MEASURES

The first relevant performance measure is the network-wide *fresh call blocking probability* $\mathbf{P}_f$, which is readily seen to be equal to the cell-specific blocking probability $\mathbf{P}_{f,b}$, $b \in \mathbb{B}$, due to symmetry. Since the fresh call arrival process is Poisson, $\mathbf{P}_{f,b}$ is equal to the fraction of time that cell b is in a state of congestion (PASTA property [224]), so that

$$\mathbf{P}_f = \mathbf{P}_{f,b} = \sum_{\mathbf{n} \in \mathbb{S} \setminus \mathbb{S}_b^\star} \pi(\mathbf{n}), \text{ for any } b \in \mathbb{B}.$$

The second performance measure of interest is the *handover failure probability* $\mathbf{P}_{h,bb'}$, i.e. the probability that a handover attempt from cell b to cell b' is blocked. Since handover requests do not follow a Poisson process, the PASTA property *cannot* be applied to determine $\mathbf{P}_{h,ij}$ (see Remark 8.1). Instead the handover failure probability is a Palm probability associated with the process counting the transitions in which a call experiences either a successful or unsuccessful handover (see e.g. [36, 209]). The probability flux of successful and failed handover attempts from cell b to cell $b' \in \mathbb{B}_b$ in system state $\mathbf{n} \in \mathbb{S}$, denoted $\widetilde{\varrho}_{h,bb'}(\mathbf{n})$ and $\widehat{\varrho}_{h,bb'}(\mathbf{n})$, respectively, are given by

$$\begin{cases} \widetilde{\varrho}_{h,bb'}(\mathbf{n}) \equiv n_b \zeta(v, R)\, 1\{\mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'} \in \mathbb{S}\}, \\ \\ \widehat{\varrho}_{h,bb'}(\mathbf{n}) \equiv n_b \zeta(v, R)\, 1\{\mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'} \notin \mathbb{S}\}. \end{cases}$$

The handover failure probability $\mathbf{P}_{h,bb'}$ is then equal to

$$\begin{aligned} \mathbf{P}_{h,bb'} &= \frac{\sum\limits_{\mathbf{n} \in \mathbb{S}} \pi(\mathbf{n})\, \widehat{\varrho}_{h,bb'}(\mathbf{n})}{\sum\limits_{\mathbf{n} \in \mathbb{S}} \pi(\mathbf{n}) \left(\widetilde{\varrho}_{h,bb'}(\mathbf{n}) + \widehat{\varrho}_{h,bb'}(\mathbf{n}) \right)} \\ &= \mathbf{L}_b^{-1} \sum_{\mathbf{n} \in \mathbb{S}} \pi(\mathbf{n})\, n_b 1\{\mathbf{n} - \mathbf{e}_b + \mathbf{e}_{b'} \notin \mathbb{S}\}, \end{aligned} \tag{8.3}$$

for $b \in \mathbb{B}$ and $b' \in \mathbb{B}_b$, where $\mathbf{L}_b \equiv \sum_{\mathbf{n} \in \mathbb{S}} \pi(\mathbf{n})\, n_b$ denotes the expected number of calls served at NodeB b. Symmetry implies that $\mathbf{P}_{h,bb'} = \mathbf{P}_h$ for all $b \in \mathbb{B}$ and $b' \in \mathbb{B}_b$, where $\mathbf{P}_h$ denotes the probability that an arbitrary handover attempt fails.

Remark 8.1 As the handover rates are state-dependent, and hence do not constitute a Poisson process, even in the absence of a resource reservation in support of handovers, the handover failure probability is generally *not* equal to the fresh call blocking probability, in contrast to the assumption made in a multitude of publications [84, 108, 165, 217, 229]. An additional rationale for the difference between handover failure and fresh call blocking probabilities that is particular for CDMA-based networks, is that a handover call releases resources in its cell of origin and thus implicitly increases the capacity of the adjacent target cell.

A third principal performance measure is the *call dropping probability* $\mathbf{P}_d$, which is defined as the probability that an admitted call departs from the system prematurely due to a rejected handover request. It is stressed that $\mathbf{P}_d$ is *not* equal to the probability that an incidental handover request is rejected. Similar to the handover failure probability, the call dropping probability is a Palm probability associated with the process counting the transitions in which a call terminates either successfully or prematurely [209]. The probability flux of successful and premature call terminations in system state $\mathbf{n} \in \mathbb{S}$, denoted $\varrho_s(\mathbf{n})$ and $\varrho_d(\mathbf{n})$, respectively, are given by

$$\begin{cases} \varrho_s(\mathbf{n}) \equiv \sum\limits_{b\in\mathbb{B}} n_b \mu, \\ \\ \varrho_d(\mathbf{n}) \equiv \sum\limits_{b\in\mathbb{B}, b'\in\mathbb{B}_b} \widehat{\varrho}_{h,bb'}(\mathbf{n}). \end{cases}$$

The call dropping probability $\mathbf{P}_d$ is then equal to

$$\begin{aligned} \mathbf{P}_d &= \frac{\sum\limits_{\mathbf{n}\in\mathbb{S}} \pi(\mathbf{n})\, \varrho_d(\mathbf{n})}{\sum\limits_{\mathbf{n}\in\mathbb{S}} \pi(\mathbf{n}) \left(\varrho_s(\mathbf{n}) + \varrho_d(\mathbf{n})\right)} \\ &= \frac{\sum\limits_{\mathbf{n}\in\mathbb{S}} \pi(\mathbf{n}) \left(\zeta(v,R) \sum\limits_{b\in\mathbb{B}, b'\in\mathbb{B}_b} n_b \mathbf{P}_{h,bb'} \right)}{\sum\limits_{\mathbf{n}\in\mathbb{S}} \pi(\mathbf{n}) \left(\sum\limits_{b\in\mathbb{B}} n_b\mu + \zeta(v,R) \sum\limits_{b\in\mathbb{B}, b'\in\mathbb{B}_b} n_b \mathbf{P}_{h,bb'} \right)} \\ &= \frac{|\mathbb{B}_1|\, \zeta(v,R)\, \mathbf{P}_h}{\mu + |\mathbb{B}_1|\, \zeta(v,R)\, \mathbf{P}_h}, \end{aligned} \tag{8.4}$$

where the second equality follows from substitution of expression (8.3), and the final equality exploits the model's symmetry in using $\mathbf{P}_{h,bb'} = \mathbf{P}_h$ for all $b \in \mathbb{B}$ and $b' \in \mathbb{B}_b$, and $|\mathbb{B}_b| = |\mathbb{B}_1|$ for all $b \in \mathbb{B}$. Observing that $|\mathbb{B}_1| \, \zeta(v, R)$ is the aggregate per call handover rate, summed over all potential destination cells, the resulting expression is not surprising as it equals the probability that an arbitrary departing call, departs as a result of a failed handover.

Remark 8.2 Expression (8.4), which relates the handover failure and call dropping probabilities, particularly when rewritten in the form

$$\mathbf{P}_d^{-1} = \left(\frac{\mu}{|\mathbb{B}_1| \, \zeta(v, R)} \right) \mathbf{P}_h^{-1} + 1,$$

may confuse the reader into thinking that $\mathbf{P}_d$ in/decreases iff $\mathbf{P}_h$ does. Consider a simple experiment where the terminal velocity v is varied, while keeping the generally velocity-dependent κ's fixed for reasons of simplicity. Numerical evaluations (carried out for various κ's and ρ's; not included) indicate that, as expected, the dropping probability *increases* in v, as the number of handover attempts per call increases and hence also the likelihood that one of them fails. The fresh call blocking and handover failure probability *decrease* in v, however, as a consequence of the reduced carried traffic load. Hence in expression (8.4) the call dropping probability $\mathbf{P}_d$ is more sensitive to v via the handover request rate $\zeta(v, R)$ (direct, autonomous effect) than via the handover failure probability $\mathbf{P}_h$ (indirect effect).

As an additional insightful performance measure, the *expected sojourn time* of a call can be determined using Little's formula (see e.g. [213]) and exploiting symmetry:

$$\mathbf{W} = \frac{\sum\limits_{\mathbf{n} \in \mathbb{S}} \pi(\mathbf{n}) \, n_1}{(1 - \mathbf{P}_f) \, \lambda(\lambda_o, R)}.$$

In general, $\mathbf{W} \leq \mu^{-1}$ must obviously hold, i.e. the expected sojourn time of a potentially prematurely terminated call cannot exceed the expected sojourn time in case all calls are allowed to terminate successfully.

Remark 8.3 In case the suggested approximations are applied, recall that the degree of mobility $\zeta(v, R)$ does not appear in the equilibrium distribution $\widehat{\pi}_{\text{PT}}(\mathbf{n})$ $(= \widehat{\pi}_{\text{BM}}(\mathbf{n}))$, so that the approximate fresh call blocking probability *only* depends on the terminal velocity through its impact on the state space. We further note that the derivation of all GOS measures follows the dynamics of the *original* Markov chain, ignoring the Markov chain modifications that underly the approximations.

8.6.3. NETWORK OPTIMISATION

For given environment parameters λ_o, μ, v, and control parameters R and RRR, the performance measures $\mathbf{P}_f$ and $\mathbf{P}_d$ can now be calculated. As the outage requirement $\mathbf{P}_o^\star$ is already met by an appropriate choice of the state space $\mathbb{S}$, the reduced objective of a network operator is to minimise the investment costs, i.e. to maximise R, such that its blocking and dropping probability targets, denoted $\mathbf{P}_f^\star$ and $\mathbf{P}_d^\star$, respectively, are met, where RRR is utilised to balance $\mathbf{P}_f$ and $\mathbf{P}_d$ appropriately. Stated formally:

$$\text{OBJECTIVE:}\quad \max\ R\ \ \text{subject to } \mathbf{P}_f \leq \mathbf{P}_f^\star \text{ and } \mathbf{P}_d \leq \mathbf{P}_d^\star.$$

An iterative procedure is implemented to achieve this objective and determine the optimal settings of the control parameters. Consisting of an outer shell to adjust cell radius R and an inner shell to adjust reservation level RRR, the procedure decreases an initially large cell radius until a reservation level can be found such that both performance requirements are satisfied. Although standard bisection search methods cannot be applied in determining the optimal cell radius, due to the fact that $\mathbf{P}_d$ is not monotonous in R (see below), the procedure can be accelerated significantly by intelligently choosing initial values for R.

8.7. NUMERICAL RESULTS

This section presents the results of a numerical study into the impact of terminal mobility on UMTS network planning. As the primary purpose of this study is to provide qualitative insight, consideration of a 2×2 network suffices, as it captures the principal features of the UMTS network that are relevant for our investigation, i.e. inter-cellular interference and terminal mobility. An additional advantage is that of reduced computational complexity. A distinction is made between three cases regarding the instruments and knowledge available to the network operator:

- the *optimal performance* is determined assuming the availability of both R and RRR as optimisation parameters, while the operator is assumed to have perfect knowledge regarding the terminal velocity in its network;

- a *suboptimal performance* is determined assuming the restricted availability of only R as an optimisation parameter, in order to indicate the system performance degradation in case an operator does not deploy a Radio Resource Reservation scheme; as above, the operator is assumed to have perfect knowledge regarding the terminal velocity in its network;
- a *sensitivity analysis* is conducted, in order to indicate the impact of a false assumption regarding terminal mobility; in this case R and RRR are optimised for a *(supposed)* terminal velocity of 3 km/h, while the system performance is evaluated for an *(actual)* terminal velocity of v which is either 3, 50 or 120 km/h; this analysis quantifies a trade-off between investment cost savings versus performance degradation.

The numerical results are presented in graphs below, depicting the (sub)optimal setting of the steering parameters R and RRR, the corresponding network investment costs, and the corresponding blocking and dropping probabilities with respective target values $\mathbf{P}_f^\star = 1\%$ and $\mathbf{P}_d^\star = 0.1\%$, all shown as a function of the nominal traffic load ρ_o (in Erlang/km^2). Given cell radius R (in km), the corresponding network investment costs are derived considering a given area that needs to be covered:

$$\begin{aligned}\text{investment costs} &= \frac{\text{network service area}}{\text{cell service area}} \times \text{cost per NodeB}\\ &= \frac{41532}{\frac{3}{2}R^2\sqrt{3}} \times \$60000,\end{aligned}$$

where 41532 km^2 is the total area of The Netherlands, used as an example, and \$60000 is the assumed cost per NodeB.

Prior to the presentation and discussion of the numerical results for the three cases defined above, some insightful results are provided in order to demonstrate the impact of the cell radius on the dropping probability. This relation turns out to be non-monotonous, which precludes a straightforward and efficient bisection search for the optimal steering parameter(s) R (and RRR).

8.7.1. THE IMPACT OF THE CELL RADIUS ON $\mathbf{P}_d$

As was argued in Section 8.3, on the one hand a denser site planning (i.e. a smaller radius R) reduces the traffic load $\rho(\lambda_o, R)$ per cell and consequently the number of calls requesting a handover; while on the other hand, it reduces the cell residence time per call, thus increasing the call-specific handover request rate $\zeta(v, R)$. The net effect depends on e.g. the nominal traffic load ρ_o, the uniform terminal velocity v, and the Radio Resource Reservation level RRR.

For a video traffic load of $\rho_o = 0.5$ Erlang/km^2, Figure 8.6 (right) shows $\mathbf{P}_d$ as a function of the cell radius R for $v \in \{3, 50, 120\}$ km/h. With regard to the two opposite effects described above, the figure illustrates that up to a so-called *break-even* cell radius $R^\star$ the former effect dominates, i.e. an increase of the cell radius and the corresponding higher traffic load induces higher aggregate handover request rates, whereas beyond $R^\star$ the latter effect dominates, i.e. the increased cell residence times outweigh the increase of the traffic load. As $R \to \infty$, the call-specific handover request rate approaches zero, while the number of calls present in the system converges to the deterministic maximum of $\lfloor \kappa_r + 3\,\kappa_a \rfloor^{-1}$. As a consequence the dropping probability converges to zero. The figure further shows the intuitively clear result that the dropping probability is greater for higher terminal velocities.

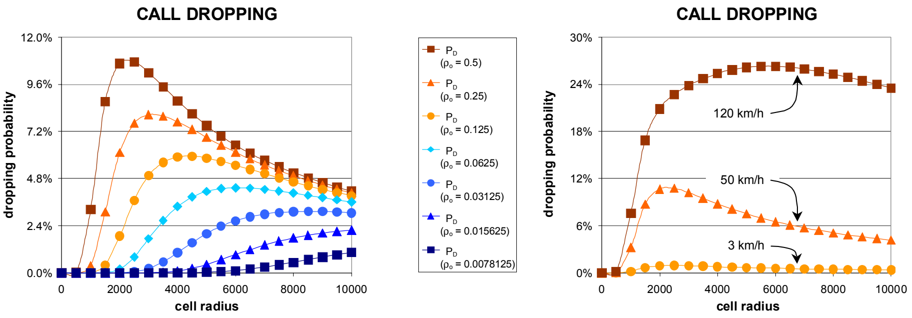


Figure 8.6 The impact of the cell radius on the dropping probability ($\mathbf{P}_d$) for different nominal video traffic loads $\rho_o \in \{0.0078125, 0.015625, 0.03125, 0.0625, 0.125, 0.25, 0.5\}$ Erlang/km^2 ($v = 50$ km/h; left) and different terminal velocities $v \in \{3, 50, 120\}$ km/h ($\rho_o = 0.5$ Erlang/km^2; right) (video service).

Figure 8.6 (left) shows $\mathbf{P}_d$ as a function of the cell radius R for $\rho_o \in \{0.0078125, 0.015625, 0.03125, 0.0625, 0.125, 0.25, 0.5\}$ Erlang/km^2 ($v = 50$ km/h, no RRR applied). While it is obvious that an increased nominal traffic load leads to higher dropping probabilities, it also appears to lower the break-even cell radius $R^\star$. This can be understood as follows. Consider e.g. the curve for $\rho_o = 0.5$ Erlang/km^2. As R

increases towards $R^\star(0.5) \approx 2500$, the dropping probability goes up due to an increase in the (offered) traffic load ρ that outweighs the effect caused by a reduction in the call-specific handover request rate $\zeta(v, R)$. Whereas initially, i.e. for low R, the *carried* traffic load grows almost in line with the *offered* traffic load, as R increases, the growth of the carried traffic load is reduced due to an increasing fresh call blocking probability. In particular, for $R > R^\star(0.5)$, the continuing reduction of the call-specific handover request rate $\zeta(v, R)$ becomes dominant and the dropping probability decreases consequently. Now consider the case with a lower nominal traffic load, e.g. $\rho_o = 0.0078125$. While the call-specific handover request rate $\zeta(v, R)$ is independent of ρ_o and thus decreases with R in the same way as it does for $\rho_o = 0.5$, the *carried* load continues to increase significantly with R well beyond $R^\star(0.5)$, in accordance with the more slowly increasing fresh call blocking probability. It is therefore not surprising that the corresponding break-even cell radius $R^\star(0.0078125) > R^\star(0.5)$, or more generally, that $R^\star(\rho_o) > R^\star(\rho_o')$ if $\rho_o < \rho_o'$.

8.7.2. OPTIMAL PERFORMANCE (R AND RRR)

For both speech and video calls, a range of nominal traffic loads ρ_o and three different terminal velocities (3, 50 and 120 km/h) with the corresponding E_b/N_o requirements, the optimal cell radius R and reservation level RRR are determined and depicted in Figures 8.7 and 8.8, along with the corresponding blocking and dropping probabilities (the same 'legend sharing' method is applied as in e.g. Chapter 3). Due to the extreme computational expenses, the numerical results for the speech service have been obtained by means of Markov chain simulation, rather than numerical calculation with the successive overrelaxation method that is used for the video service. 95% confidence intervals have been determined for these simulation results, with a relative precision not worse than 5%. The same optimisation procedure has been applied to both the speech and video services.

Aside from the unsurprising result that the optimal cell radius is decreasing in both the traffic load ρ_o and the terminal velocity v, the results demonstrate that the deployment of a radio resource reservation scheme can indeed be effectively utilised to reduce call dropping, as the optimal setting of RRR is greater than 0 for both cases with a significant degree of mobility ($v \in \{50, 120\}$ km/h) and considering sufficiently high (yet realistic) traffic loads. Note that for all ρ_o and v the blocking and dropping requirements are indeed met as intended, but that the requirements are generally not

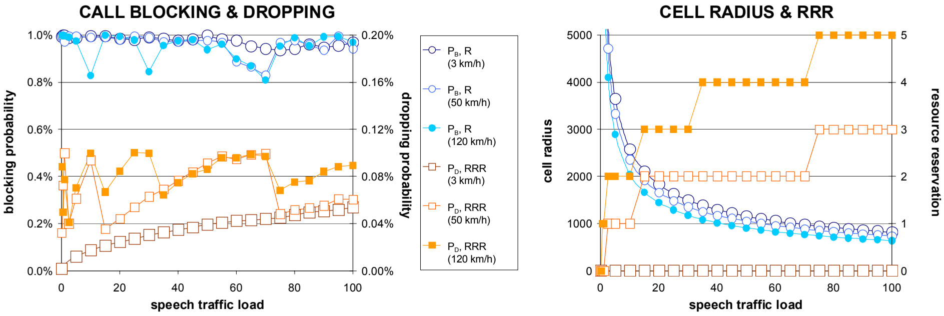


Figure 8.7 Optimal case: call blocking ($\mathbf{P}_f$) and dropping ($\mathbf{P}_d$) probabilities for a range of nominal traffic loads (left); settings of cell radius R and Radio Resource Reservation level RRR for a range of nominal traffic loads (right) (speech service).

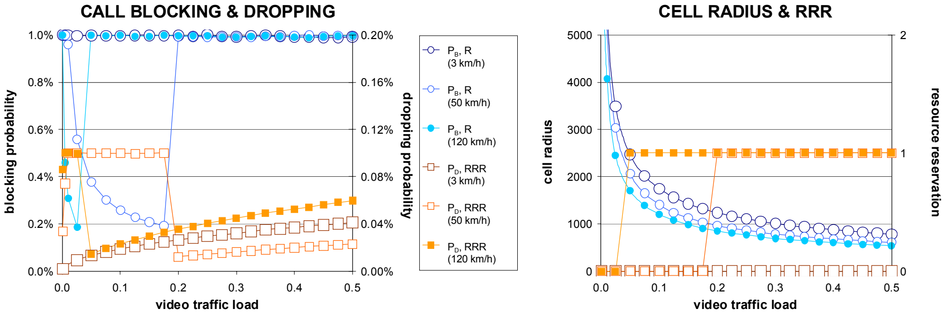


Figure 8.8 Optimal case: call blocking ($\mathbf{P}_f$) and dropping ($\mathbf{P}_d$) probabilities for a range of nominal traffic loads (left); settings of cell radius R and Radio Resource Reservation level RRR for a range of nominal traffic loads (right) (video service).

met with equality. Although this observation may seem to indicate that there is still some room for improvement, i.e. a larger R, this is not the case, primarily due to the fact that one of the steering parameters (RRR) is restricted to take on only positive integer values.

Consider the illustrative case of video calls with $v = 50$ km/h. For $\rho_o \in [0, 0.2)$, the optimal cell radius is determined by $\mathbf{P}_d^\star$ only, while the degree of mobility is still not sufficient to justify the deployment of Radio Resource Reservation. As the cell radius decreases to satisfy $\mathbf{P}_d \leq \mathbf{P}_d^\star$, $\mathbf{P}_f$ falls exponentially due to the reduced video traffic load per cell. At $\rho_o = 0.2$, the degree of mobility has become sufficiently large, due to relatively small cell sizes in combination with the high video traffic load, to justify $RRR = 1$. For video calls with $v = 50$ km/h, $RRR = 1$ corresponds to a reservation of 20% of a cell's pole capacity of $\lfloor 1/\kappa_r \rfloor = 5$. As a consequence of this reservation, $\mathbf{P}_d$ decreases drastically, while $\mathbf{P}_f^\star$ is now the bottleneck restriction. Note the correspondence between the RRR increments in Figure 8.8 (right) and the sharp decrease (increase) of $\mathbf{P}_d$ ($\mathbf{P}_f$) in Figure 8.8 (left). As ρ_o increases further, the cells continue to shrink, now in order to satisfy $\mathbf{P}_f \leq \mathbf{P}_f^\star$, while $\mathbf{P}_d$ gradually increases. Even if ρ_o becomes extremely large, requiring a cell radius of less than 10 meters, an upgrade of the radio resource reservation to $RRR = 2$ is not justified (for $v = 50$ km/h).

Figure 8.9 shows the investment costs for the speech (left) and video (right) services, corresponding with the optimal cell radii of Figures 8.7 and 8.8, respectively. The results show that although the optimal cell radius seems to become decreasingly sensitive to the traffic load, the investment costs are approximately linear in ρ_o, which is intuitively supported by the following argument. In the specific case without terminal mobility ($v = \mathbf{P}_d = 0$), it is readily understood that the network is optimised if the offered traffic load $\rho^\star$ per cell is established that meets $\mathbf{P}_f = \mathbf{P}_f^\star$ precisely. The optimal cell radius is then directly given by $R^\star = \sqrt{\rho^\star / \left(\frac{3}{2}\rho_o\sqrt{3}\right)}$, so that the corresponding investment costs are equal to $41532 \times \$60000 \,/\, \left(\frac{3}{2}\,(R^\star)^2\sqrt{3}\right) = 41532 \times \$60000 \times \rho_o/\rho^\star$, which is indeed linear in ρ_o. This simple best-case analysis presumes the blocking probability to be the bottleneck performance measure, which is not necessarily the case for networks *with* terminal mobility, as demonstrated above. However, it is apparent that although $v > 0$ can have a significant influence on required investment costs, the linearity in the nominal traffic load is approximately preserved.

As an illustrative case of the velocity-induced difference in the required investment costs, a speech traffic load of $\rho_o = 50$ Erlang/km^2 requires a perfectly realistic cell radius of $R = 1156$ ($v = 3$), $R = 1049$ ($v = 50$), or $R = 909$ ($v = 120$) meters to meet the GOS targets, corresponding with a calculated investment cost of $\$718 \cdot 10^6$, $\$872 \cdot 10^6$, or $\$1161 \cdot 10^6$, respectively. Hence the investment cost induced by a terminal velocity of 120 km/h are more than 60% higher than that induced by a terminal velocity of 3 km/h, while it is increasing in ρ_o. Observe even more extreme differences in Figure 8.9 (right) for the considered range of video traffic loads.

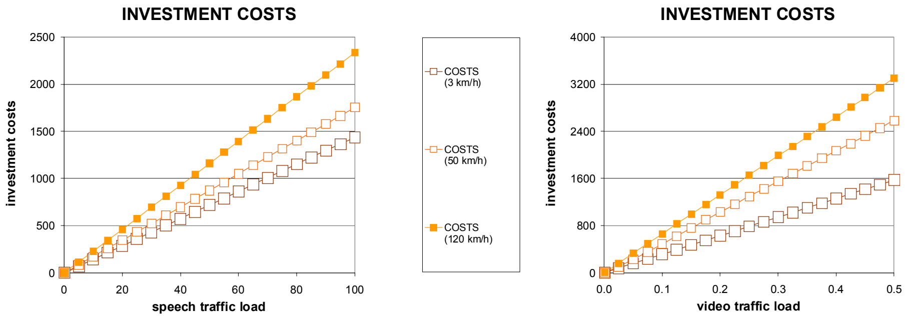


Figure 8.9 Optimal case: investment costs corresponding to the optimal cell radii for a range of nominal traffic loads for the speech (left) and video service (right).

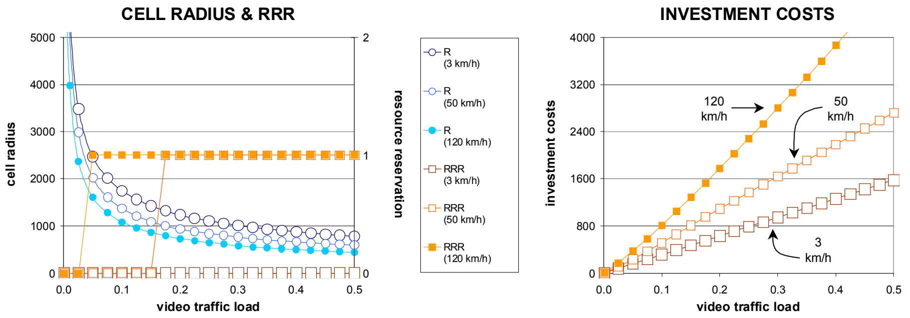


Figure 8.10 Optimal case (approximation): settings of Radio Resource Reservation level *RRR*, cell radius *R* (left) and the corresponding investment costs (right) for a range of nominal video traffic loads.

APPROXIMATIONS

As an indication of the performance of the product-form approximations outlined in Section 8.6.1, Figure 8.10 depicts the ‘optimal’ cell radius and Radio Resource Reservation level versus the video traffic load (left chart) and the associated investment costs (right chart). In each step of the optimisation procedure, the performance measures are derived from the product-form equilibrium distribution of the modified Markov chain. Although the general trends are equivalent to the corresponding ‘true’ performance depicted in Figures 8.8 (right) and 8.9 (right), observe that, particularly for higher terminal velocities, the obtained cell radii are *smaller* than the optimised values, leading to a potentially significant overinvestment. This is a consequence of the fact that the model approximations made to obtain the product-form equilibrium distribution, effectively raise the carried load of the system by retaining otherwise dropped calls, leading to an overestimate of the Grade of Service measures and, ultimately smaller cells. It is stressed that in a growing market, some degree of network overdimensioning is desired, so that application of the proposed approximations may not be harmful. As expected, note that for a low degree of terminal mobility ($v = 3$) the product-form approximation appears to be excellent.

8.7.3. SUBOPTIMAL PERFORMANCE (R ONLY)

Consider the suboptimal case allowing the system optimisation over the cell radius R only. Figure 8.11 (right) depicts the obtained suboptimal value of R for the same range of *video* traffic loads as above, along with the corresponding investment costs, while Figure 8.11 (left) depicts the system performance in terms of the call blocking and dropping. The first observation is that the cell radius required to meet the blocking and dropping probability targets is smaller in the considered suboptimal case where no RRR is deployed, and hence the investment costs are higher. Note further that for $v \in \{50, 120\}$ km/h, $\mathbf{P}_d$ appears to be the bottleneck measure in the optimisation procedure, since the resulting $\mathbf{P}_f$ are well below the target value of 1%. In contrast, for $v = 3$ km/h the degree of mobility is so low that $\mathbf{P}_d$ never even approaches $\mathbf{P}_d^\star$, and the performance optimisation is steered solely by $\mathbf{P}_f$. In fact, the performance and investment costs for $v = 3$ km/h are identical to those obtained above for the case *with* the RRR option, as the optimal setting of RRR was found to be 0 anyway.

To evaluate the gain from the deployment of an appropriate RRR scheme, compare the required investment costs for the video service in Figures 8.9 (right) and 8.11 (right). Obviously, for the case of $v = 3$ km/h, there is no gain whatsoever, as $RRR = 0$ was found to be optimal. For a moderate velocity of $v = 50$ the gain becomes slightly more apparent (for sufficiently large loads), establishing investment cost savings that are increasing in ρ_o within its depicted range, reaching a maximum of about 8.98% for $\rho_o = 0.5$ Erlang/km^2. Most significantly, for the case of $v = 120$ km/h optimal deployment of Radio Resource Reservation induces cost savings of as much as 22.34%, achieved for $\rho_o = 0.5$, while the gains appears to increase with ρ_o.

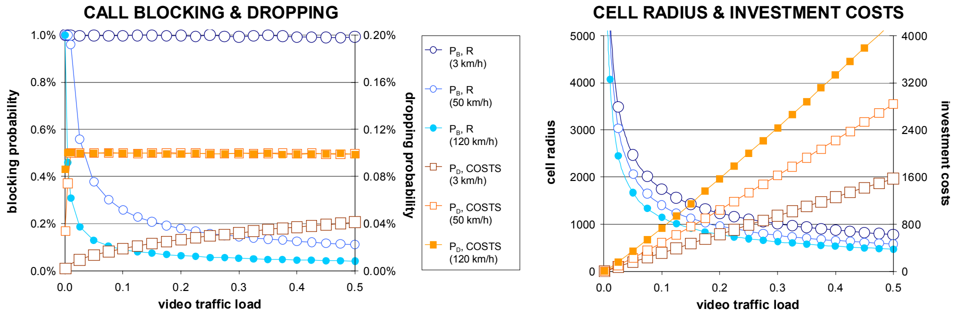


Figure 8.11 Suboptimal case: setting of the cell radius R and the corresponding investment costs, call blocking ($\mathbf{P}_f$) and dropping ($\mathbf{P}_d$) probabilities for a range of nominal traffic loads (video service).

The trends that are obtained for the speech service are equivalent to those discussed for the video service above (not shown).

8.7.4. SENSITIVITY ANALYSIS

Figure 8.12 shows the results of a sensitivity analysis that has been carried out in order to investigate the impact of a false assumption regarding terminal mobility on the system performance (*video* service). The sensitivity analysis is based on the operator's assumption that $v = 3$ km/h, which is thus used as a given parameter in the optimisation of R and RRR. The right chart depicts the obtained cell radius R and the corresponding investment costs, while $RRR = 0$ for all $\rho_o \in [0, 0.5]$, all of which are identical to those given in the previous numerical subsections. The left chart then shows the corresponding $\mathbf{P}_f$ and $\mathbf{P}_d$ for $v = 3$, 50 and 120 km/h. Note that for the case where the operator *correctly* supposes that $v = 3$ km/h, the GOS measures obviously meet their target values (as shown before).

It is clear that $v \in \{50, 120\}$ km/h provides the more relevant cases in the sensitivity analysis. The depicted GOS measures indicate the performance cost that is induced by the false assumption regarding the terminal velocity, and establish a trade-off with respect to the investment cost savings that can be deduced from Figure 8.9 (right). Within the given range of traffic loads, the cost reduction may be as significant as 39.29% ($v = 50$ km/h, $\rho_o = 0.325$) or 52.48% ($v = 120$ km/h, $\rho_o = 0.325$),

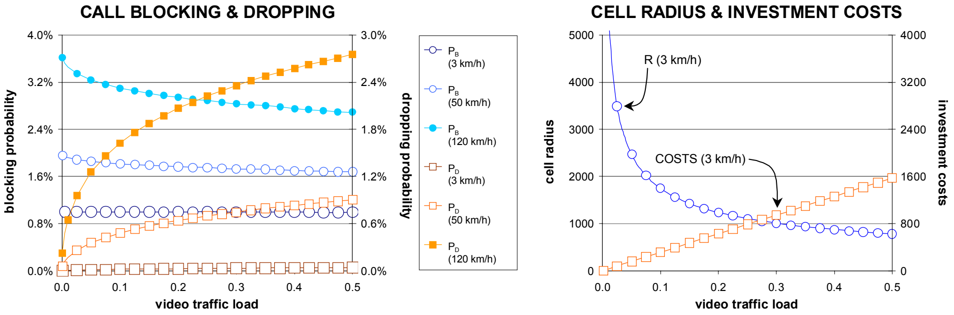


Figure 8.12 Sensitivity analysis: setting of the cell radius R and the corresponding investment costs, call blocking ($\mathbf{P}_f$) and dropping ($\mathbf{P}_d$) probabilities for a range of nominal traffic loads (video service).

while the corresponding performance degradation raises $\mathbf{P}_f$ to 1.72% $> \mathbf{P}_f^\star$ and $\mathbf{P}_d$ to 0.77% $> \mathbf{P}_d^\star$, for $v = 50$ km/h, while it raises $\mathbf{P}_f$ to 2.82% $> \mathbf{P}_f^\star$ and $\mathbf{P}_d$ to 2.42% $\gg \mathbf{P}_d^\star$, for $v = 120$ km/h. In particular, if the actual terminal velocity differs greatly from the supposed terminal velocity, the apparent impact on the dropping probability is most significant and, in relative terms, rather dramatic. It is therefore concluded that it is important for an operator to predict the terminal velocity in the different areas with reasonable accuracy.

A final observation that is made from Figure 8.12 (left) is that for $v \in \{50, 120\}$ km/h, $\mathbf{P}_f$ appears to be decreasing in the *offered* traffic load ρ_o. This seemingly counterintuitive trend is readily explained by the increase of $\mathbf{P}_d$, due to a higher degree of mobility, which reduces the effectively *carried* traffic load.

The trends that are obtained for the speech service are equivalent to those discussed for the video service above, although the effect of a significant underestimation of the terminal velocity on the established GOS performance appears to be even more dramatic (not shown).

8.8. CONCLUDING REMARKS

We have presented an investigation into the impact of terminal mobility on UMTS radio network planning, which has been argued to originate from the more severe E_b/N_o requirements that apply in case of higher velocities, and the greater Radio

Resource Reservation that is required in case of higher terminal velocities, in order to protect on-going calls from being dropped in a handover attempt.

The considered objective has been to minimise network investment costs (or: maximise the cell radius) as a function of the traffic load per km^2 and the terminal velocity, using Radio Resource Reservation to affect the trade-off between call blocking and dropping. An $M \times M$ cellular torus-shaped UMTS network model has been used, whose CDMA-specific soft capacity has been converted into hard capacity by posing a minimum QOS requirement for active calls, allowing Markov chain analysis. Aside from the brute-force numerical derivation of the Markov chain's equilibrium distribution using e.g. a successive overrelaxation algorithm, a model modification has been suggested which leads to a product-form approximation. The principal strength of the presented model and approach lies therein that it is simple enough to allow a computationally relatively inexpensive performance evaluation and optimisation, yet sufficiently realistic to suggest rough guidelines for cell dimensioning and RRR deployment, and to provide valuable qualitative insight for network planning purposes.

The primary conclusions from the included numerical experiments are fourfold. With regard to the derivation of the Markov chain's equilibrium distribution, an illustrative case study indicates that the product-form approximation works fine for low terminal velocities, but may lead to significant network overdimensioning if the degree of mobility is high. Concerning the actual research objectives of this investigation, we conclude that the impact of terminal velocity on the optimal cell radius and, as a derivative, on the investment costs can be very significant (potentially even well beyond a factor 2!). Further, the deployment of a Radio Resource Reservation scheme can indeed be effectively utilised to reduce call dropping and investment costs. Lastly, planning a UMTS network using inaccurate estimates of terminal velocity potentially leads to inacceptable blocking and, in particular, dropping probabilities.

CHAPTER 9

PERFORMANCE ANALYSIS OF WIRELESS LOCAL AREA NETWORKS

WIRELESS Local Area Networks are expected to play an important role in future everyday's communication, not only in the private domain but also for public use. In particular, they may fulfill the need for an additional public wireless access solution for data services in hot spots (e.g. train stations, airports, etc.), besides the access provided by mobile cellular networks such as GSM/GPRS and UMTS [109]. WLANs provide an interesting possibility to offer additional capacity and higher bandwidths to end-users without sacrificing the inherently scarce and expensive capacity of cellular networks. However, critical factors for a successful introduction are security and performance, which applies in particular to deployment of WLANs in the public environment.

WLAN performance is largely determined by the maximum data rate at the physical layer and the MAC (Medium Access Control) layer protocols defined by the IEEE 802.11 standards [110, 111, 189]. The most widely employed WLAN MAC protocol is the Distributed Coordination Function (DCF). The DCF is a random access scheme based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), which uses random backoffs in order to manage packet retransmissions in case of a destructive collision. If the DCF is used in its BASIC access mode, the aggregate WLAN throughput decreases drastically for a larger number of active stations, due to a rapidly increasing number of collisions. The occurrence of collisions is particularly significant in cases with so-called hidden stations, i.e. when stations cannot detect each other's activity simply by sensing the medium. In order to overcome this throughput degradation the Request-To-Send/Clear-To-Send (RTS/CTS) mechanism has been standardised, where a station sends a small control packet in order to reserve the channel for transmission of a data packet. Aside from the DCF protocol, the IEEE 802.11 standard also defines an optional, centralised MAC protocol called Point Coordination Function (PCF). In the PCF scheme a central node polls the stations to access the shared medium, thus eliminating the need for contention and enabling the support of delay-sensitive services. The remainder of this chapter focuses on the DCF MAC protocol.

The IEEE 802.11 WLAN can operate in infrastructure mode or in ad hoc mode, see Figure 9.1. In ad hoc mode all stations can transmit packets directly to other stations that are within the sending range (Basic Service Set (BSS)). In the infrastructure mode an Access Point is present to link the stations in a BSS. An AP may be connected to a distribution system (e.g. a wired LAN) via which stations are linked to other APs or e.g. a remote server can be reached.

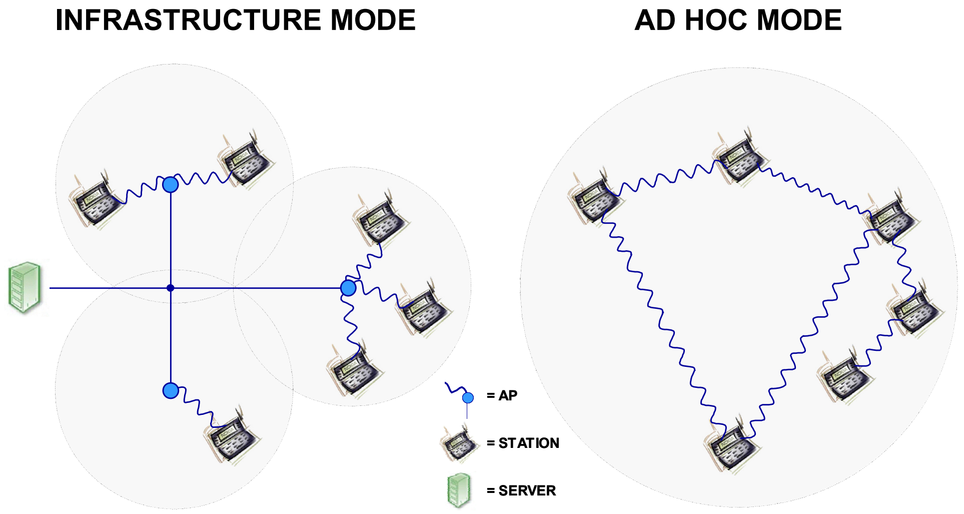


Figure 9.1 The IEEE 802.11 standard features two WLAN operating modes: the infrastructure mode (left) and the ad hoc mode (right).

The outline of this chapter is as follows. Section 9.1 reviews the relevant literature, followed by a statement of the chapter's contribution in Section 9.2. In Section 9.3, the IEEE 802.11 DCF MAC protocol is explained in more detail. Section 9.4 describes the system, traffic and capture models underlying the analytical performance study, which relies on a separation of time scales, as presented in Section 9.5. In Section 9.6 we present an extensive numerical study in order to validate the accuracy of our analytical model (by comparison with simulation) and to illustrate the impact of various model parameters on the system performance. Finally, the principal conclusions of our investigation as well as some topics for further research are outlined in Section 9.7.

9.1. LITERATURE

A number of papers have studied the throughput performance of IEEE 802.11 DCF MAC for both the BASIC and RTS/CTS access modes. Several of them are based on simulation, see e.g. [222]. Other papers use analytical models, but with simplifying assumptions about the DCF MAC layer operations and/or the traffic conditions in order to enable mathematical analysis. In particular, a strongly simplified backoff mechanism has been used in [40, 48], while [103, 230] assume Poisson sources generating fixed size data packets. A more detailed mathematical performance model of DCF has been developed and analysed by Bianchi [26] that was slightly improved by Wu et al. [225]. The key approximation made in these papers enabling a relatively simple Markov chain analysis is the assumption of independent transmissions by different flows, as well as constant and independent collision probabilities, regardless of the number of erroneous transmissions experienced. Comparison with simulation shows that the analytical results are generally accurate. [26, 225] both assume a constant number of persistently active stations and a simplified physical layer model. In [96, 97] a more realistic physical layer model is studied in order to assess the impact of *packet capture* on the aggregate system throughput, where a transferred data packet may survive a collision with concurrently transferred data packets. To the author's knowledge only one paper exists, [83], which considers flow transfer times in a WLAN with non-persistent flows. Using the DCF performance model and results of [26], the authors construct a continuous-time Markov chain describing the system dynamics when the number of active stations varies in time. The steady-state distribution of this Markov chain is *numerically* solved from the balance equations and yields approximations for the expected throughput and flow transfer time.

9.2. CONTRIBUTION

In this chapter we extend the work of [26, 225] in two directions. First, we further elaborate on Bianchi's packet level model [26] by integrating the various modelling enhancements on the physical and MAC layer proposed by other authors [96, 97, 225] into a single DCF performance model, still allowing analytical treatment. Our second extension covers the practical situation that the number of active stations is not constant (as in e.g. [26, 225]) but varies in time due to the random user behaviour given by the initiation and completion of data flows. In order to enable mathematical analysis of flow throughputs and transfer times in the system with the extensions mentioned above we propose an integrated packet/flow level modelling approach along similar lines as e.g. [139]. In particular, from the flow level point of view, the WLAN is considered as a queueing system with Poisson flow arrivals and a Generalised Processor Sharing service discipline, which reflects the IEEE 802.11 DCF MAC design principle of distributing the transmission capacity fairly among the active flows. The rate at which the flows are served depends on the number of flows simultaneously present in the system (i.e. the number of active stations). These service rates are obtained from the analysis of our extended packet level model describing the behaviour of the DCF in detail for the situation with a fixed number of persistently active stations. The resulting GPS queueing model with state-dependent service rates is analytically tractable (see e.g. [54]) and yields closed-form expressions for the (conditional) expected flow transfer time (of a flow of given size). Our modelling approach provides also some important, more general insights in the essential WLAN performance characteristics. In particular, the well-known insensitivity property of the (G)PS model implies that the expected transfer times are independent of the flow size distribution, apart from its mean, while in addition, the conditional expected flow transfer time is linear in the flow size. These attractive properties and the accuracy of our analytical performance results are validated by simulations, revealing an excellent fit. Summarising, the principal contributions of the present chapter are the inclusion of an enhanced DCF and physical layer model that remains analytically solvable, and the recognition that the resulting flow level model is a tractable GPS queue, which opens the possibility for additional performance analysis of e.g. Call Admission Control.

9.3. DISTRIBUTED COORDINATION FUNCTION

The DCF is the most widely employed IEEE 802.11 MAC layer protocol [110]. It is based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) and is characterised by random backoffs that are sampled from an exponentially increasing contention window size in order to manage retransmissions upon destructive collisions. The DCF operates in either the BASIC or the RTS/CTS access mode, which are explained in more detail below.

9.3.1. BASIC ACCESS MODE

Figure 9.2 illustrates the principle of the BASIC access scheme. When a station wants to transmit a data packet, it first senses the medium to determine whether or not the channel is already in use by another station (*physical carrier sensing*). If the channel is sensed idle for a contiguous period of time called DIFS (Distributed InterFrame Space), the considered station transmits its packet. In case the channel is sensed busy, the station must wait until it becomes idle again and subsequently remains idle for a DIFS period, after which it has to wait another randomly sampled number of time slots before it is permitted to transmit its data packet. This *backoff* period is sampled from a discrete uniform distribution on $\{0, \cdots, cw_r - 1\}$, with cw_r the contention after r failed packet transfer attempts (cw_0 is the initial contention window size). The backoff counter is decremented from its initially sampled value until the packet is transferred when the counter reaches zero, unless it is temporarily 'frozen' in case the channel is sensed busy before the backoff counter reaches zero. In the latter case the station continues decrementing its backoff counter once the medium is sensed idle for at least a DIFS period. It is noted that the idea behind the random backoff procedure is to reduce the probability of *collisions*, which occur either when the backoff counters of multiple stations reach zero simultaneously, or in case a so-called hidden station fails to freeze its backoff counter when it cannot sense another station's transmission. In a collision only the strongest signal among multiple concurrent transmissions has a chance of successful *capture* by the intended receiver.

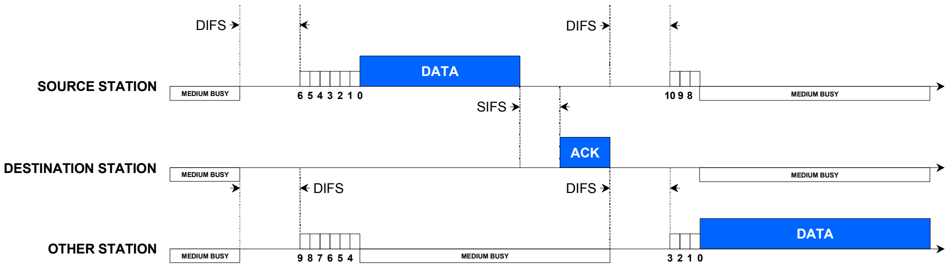


Figure 9.2 Illustration of the BASIC access mode of the Distributed Coordination Function of the IEEE 802.11 MAC layer.

If the destination station successfully captures the transmitted data packet, it responds by sending an ACK (ACKnowledgment message) after a SIFS (Short InterFrame Space) time period. A SIFS is shorter than a DIFS in order to give the ACK preference over data packet transmissions by other stations, while it is sufficiently long to allow the stations involved in the considered transfer to switch between transmission and reception mode. If the source station fails to receive the ACK within a predefined time-out period, the contention window size is adjusted according to the following expression:

$$cw_r = \begin{cases} 2^r\left(cw_{\min}+1\right), & 0 \le r \le r^\star, \\ \\ 2^{r^\star}\left(cw_{\min}+1\right), & r^\star \le r \le r_{\max}, \end{cases} \tag{9.1}$$

upon which the data packet transfer is reattempted. Here $r_{\max}$ is the maximum number of retransmissions for a given data packet and r^* is the maximum number of times that the contention window is doubled after a failed transfer attempt. Once the data packet is successfully transferred, the contention window size is reset to cw_0 and the entire procedure is repeated to transfer subsequent data packets. If an unfortunate data packet is still not successfully transferred after $r_{\max}$ retransmissions, the MAC layer gives up. It is then up to higher-layer protocols (e.g. UDP (User Datagram Protocol) or TCP) whether the packet is discarded or once again offered to the MAC layer for transmission. The interaction between these protocol layers is not explicitly investigated in this chapter.

9.3.2. RTS/CTS ACCESS MODE

Figure 9.3 illustrates the operation of the RTS/CTS mode, which features the same backoff procedure as the BASIC access mode. The principal difference between the two access modes is that the RTS/CTS access mode assumes a four- rather than a two-way handshake. Once a station is allowed to transmit, it first sends a small RTS (Request To Send) frame to the intended receiver. If the destination station properly receives the RTS frame, it responds with a CTS (Clear To Send) frame. Upon receipt of the CTS frame the source station transmits the actual data packet, which is subsequently acknowledged by the destination station. All of these frames are separated by a time period of length SIFS so that other stations cannot intervene in the sequence of transmissions.

The advantages of the RTS/CTS access mode over the BASIC access mode are twofold. First it is an efficient way to reduce the impact of a destructive collision as it is detected when the source station fails to receive a CTS frame. Since the RTS frame is typically significantly smaller than a data packet, the potential resource inefficiency incurred by a collision is smaller. A second advantage of the RTS/CTS

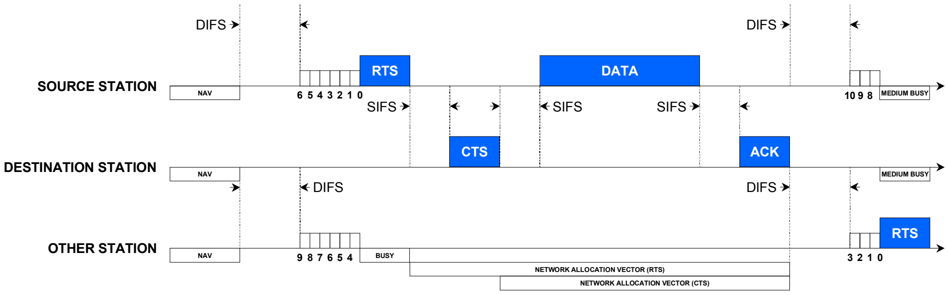


Figure 9.3 Illustration of the RTS/CTS access mode of the Distributed Coordination Function of the IEEE 802.11 MAC layer.

scheme is that it reduces the hidden station problem: even if a hidden station cannot hear the RTS frame, it may be able to hear the CTS frame, both of which contain a *duration field* that indicates the total transmission time up to and including the ACK frame. This information is used to set the station's so-called NAV (Network Allocation Vector) so that it is aware that the medium is busy, even if it cannot sense the actual transmissions directly *(virtual carrier sensing)*. The drawback of the RTS/CTS access mode is the additional overhead involved, which induces a lower effective channel utilisation for small data packets or a small number of users, i.e. when the collision probability is small.

9.4. MODEL

This section sets the framework for the presented performance analysis by describing the system, capture and traffic models in generic terms. Concrete parameter settings are specified in Table 9.1 in Section 9.6.

9.4.1. SYSTEM MODEL

We consider a single BSS with stations contending for the shared WLAN radio access medium. The fixed channel rate of the medium is denoted $r_{\text{WLAN}} \in \{1, 2, 5.5, 11\} \cdot 10^3$ kbits/s, while the physical layer preamble (required for synchronisation purposes) and header are always transmitted at a fixed rate of 1 Mbits/s to ensure compatibility between the IEEE 802.11 and IEEE 802.11B standards. The applied DCF is considered in both BASIC and RTS/CTS access mode. The DCF operations at the MAC layer are modelled in significant detail in the Markov chain that is taken from [26, 225] and specified in Section 9.5. The DCF model includes DIFS and SIFS timers, MAC layer acknowledgements, an exponentially increasing contention window, and a randomly sampled backoff counter that is decremented towards a packet transfer attempt and potentially 'frozen' if the shared medium is sensed busy. Furthermore, we integrate a more realistic physical layer model into the setting of [26, 225] by taking into account the possibility of capture in case of concurrent packet transfers. The considered capture models are discussed below after a specification of the traffic model.

9.4.2. TRAFFIC MODEL

The considered WLAN serves stations which generate data flows according to a Poisson process with rate λ. Data flows are assumed to be transfers of files with generally distributed sizes. The mean file size is denoted $1/\mu$ (in kbits). Each file is segmented into packets of a given size (with a final packet containing the flow's remainder) which are processed at the WLAN's MAC layer. The data service traffic load is denoted $\rho \equiv \lambda / (\mu r_{\text{WLAN}})$. A CAC scheme is deployed to limit the number of contending data flows to $n_{\max}$ and thus ensure system stability and provide some minimum QOS. Although strictly speaking the inclusion of CAC in the IEEE 802.11 standard family is still in preparation by task group TG-E [112], it is noted that even in the IEEE 802.11/B products some maximum exists on the number of stations that can associate with an AP.

9.4.3. CAPTURE MODELS

We apply two distinct capture models, denoted CM-1 and CM-2, which specify the likelihood that transferred data packets survive a collision with concurrently transferred packets [96, 97, 140, 221]. The common assumption underlying both capture models is that in a collision of multiple packets, *only* the one with the strongest signal has a chance of successful capture [90, 95]. In our analytical performance evaluation model, the effects of capture appear in the form of the capture functions $\mathbf{P}_s(k)$ and $\mathbf{P}_s^\star(k)$ for $k \geq 1$. The former function is defined as the probability that the *strongest* data packet among k concurrently transferred packets is successfully captured, while the latter function denotes the probability that a *tagged* data packet is successfully captured in a simultaneous transfer with $k-1$ *other* data packets. Both illustrative capture models specified below have the desired property that both $\mathbf{P}_s(k)$ and $\mathbf{P}_s^\star(k)$ are non-increasing in k.

Capture model CM-1 assumes that a packet transfer is successful if and only if there are no concurrent transfers from other flows. Expressed in the above-defined capture functions:

$$\mathbf{P}_s(k) = \mathbf{P}_s^\star(k) = \begin{cases} 0 & \text{if } k > 1, \\ \\ 1 & \text{if } k = 1. \end{cases}$$

Capture model CM-1 is the most basic option imaginable and is implicitly applied in e.g. [26, 225].

Capture model CM-2 is based on [96, 97] and assumes that signals in a collision have some uniform local mean received power $\overline{p}$, determined by attenuation and shadowing effects, which is established when all stations are located at similar distances from their intended receiver(s). The instantaneous received signal powers are independent and exponentially distributed around this mean, which is a direct consequence of an assumption of Rayleigh fading (see e.g. [141]).

Under these assumptions the capture function $\mathbf{P}_s(k)$ is equal to the probability that the carrier-to-interference ratio of the strongest signal in a collision ensemble of k signals exceeds the threshold $\Gamma^\star$ required for successful capture, i.e.

$$\mathbf{P}_s(k) = \int\limits_{p_1=0}^{\infty} \cdots \int\limits_{p_k=0}^{\infty} \left(\prod_{i=1}^{k} \varphi_{\overline{p}}(p_i) \right) 1\left\{\Gamma_{\max}(p_1, \cdots, p_k) \geq \Gamma^\star\right\} dp_1 \cdots dp_k,$$

where the p_i's, $i = 1, \cdots, k$, denote the instantaneous received signal powers of all signals involved in the collision, $\varphi_{\overline{p}}$ is the exponential PDF with mean $\overline{p}$, $1\{\cdot\}$ is the indicator function, and $\Gamma_{\max}(p_1, \cdots, p_k)$ is defined as the carrier-to-interference ratio of the strongest signal in the collision ensemble:

$$\Gamma_{\max}(p_1, \cdots, p_k) \equiv \frac{\max_i p_i}{\sum\limits_i p_i - \max_i p_i}$$

The independence of $\mathbf{P}_s\left(k\right)$ with respect to the local mean received power $\overline{p}$ follows from a substitution of $q_i \equiv p_i/\overline{p}$ in the above integral. The capture function is readily evaluated analytically, observing that

$$\begin{aligned}
&\left\{\left(p_1,\cdots,p_k\right):\Gamma_{\max}\left(p_1,\cdots,p_k\right)\geq\Gamma^\star\right\}= \\
&\qquad=\bigcup_{i=1}^{k}\left\{\left(p_1,\cdots,p_k\right):\left(1+\Gamma^\star\right)p_i\geq\Gamma^\star\sum_{j=1}^{k}p_j\right\}.
\end{aligned}$$

For the case of $\Gamma^\star \geq 1$ the sets $\left\{\left(p_1,\cdots,p_k\right):\left(1+\Gamma^\star\right)p_i\geq\Gamma^\star\sum_{i=1}^{k}p_i\right\}$ are disjoint, so that (cf. [60])

$$\begin{aligned}
\mathbf{P}_s(k) &= k\int\limits_{p_1=0}^{\infty}\cdots\int\limits_{p_k=\Gamma^\star\left(p_1+\cdots+p_{k-1}\right)}^{\infty}\left(\frac{1}{\overline{p}}\right)^k\exp\left(-\sum_{i=1}^{k}\frac{p_i}{\overline{p}}\right)dp_1\cdots dp_k \\
&= k\int\limits_{p_1=0}^{\infty}\cdots\int\limits_{p_{k-1}=0}^{\infty}\left(\frac{1}{\overline{p}}\right)^{k-1}\exp\left(-\frac{\left(1+\Gamma^\star\right)}{\overline{p}}\sum_{i=1}^{k-1}p_i\right)dp_1\cdots dp_{k-1} \\
&= \frac{k}{\left(1+\Gamma^\star\right)^{k-1}}, \qquad (9.2)
\end{aligned}$$

using the fact that the integrated probability mass of a $(k-1)$-dimensional joint exponential distribution with uniform parameter $\left(1+\Gamma^\star\right)/\overline{p}>0$ is equal to 1. Observe that the resulting expression does indeed not depend on $\bar{p}$. In the alternate case that $\Gamma^\star \leq 1$ the sets $\left\{\left(p_1,\cdots,p_k\right):\left(1+\Gamma^\star\right)p_i\geq\Gamma^\star\sum_{i=1}^{k}p_i\right\}$ are not disjoint, so that, although it is still straightforward, evaluation of $\mathbf{P}_s(k)$ requires substantial and careful bookkeeping. Another case we have worked out explicitly is that for $\Gamma^\star\in\left[\frac{1}{2},1\right]$, for which the capture function is equal to

$$\mathbf{P}_s(k)=\frac{k}{\left(1+\Gamma^\star\right)^{k-1}}-\frac{1}{2}k\left(k-1\right)\left(\frac{1-\Gamma^\star}{1+\Gamma^\star}\right)^{k-1}.$$

Since all signal powers are assumed to be independent and identically distributed, the probability that a tagged signal is the strongest one in a collision ensemble of k

signals is equal to $1/k$, so that capture function $\mathbf{P}_s^\star(k)$ for CM-2 is given by

$$\mathbf{P}_s^\star(k) = \frac{1}{k}\mathbf{P}_s(k).$$

In case $\Gamma^\star \geq 1$, the resulting expression for $\mathbf{P}_s^\star(k)$ (using (9.2)) reflects a form of independence in the sense that $\mathbf{P}_s^\star(k) = (\mathbf{P}_s^\star(2))^{k-1}$, i.e. the probability that a tagged signal is sufficiently stronger than the sum of $k-1$ interfering signals is equal to the probability that the tagged signal is sufficiently stronger in each of $k-1$ pairwise comparisons with the individual interfering signals.

The carrier-to-interference ratio requirement is given by $\Gamma^\star \equiv z_0 g(S_f)$. Here z_0 denotes the required energy-per-bit to interference-plus-noise density ratio, which typically lies somewhere in the range $6-24$ dB. Assuming rectangular chip pulses at the receiver, the inverse processing gain $g(S_f) \equiv 2/(3S_f)$ is a function of the spreading factor S_f. S_f is equal to 11 for $r_{\text{WLAN}} \in \{1, 2\} \cdot 10^3$ kbits/s (IEEE 802.11: Binary/Quadrature Phase Shift Keying), resulting in $\Gamma^\star \in [0.2413, 15.2236]$. A spreading factor of $S_f = 8$ is used for $r_{\text{WLAN}} \in \{5.5, 11\} \cdot 10^3$ kbits/s (IEEE 802.11B: Complementary Code Keying), which leads to $\Gamma^\star \in [0.3318, 20.9324]$.

9.4.4. PERFORMANCE MEASURES

The system level performance is assessed in terms of the *aggregate data throughput*, while the data QOS is expressed by the *(conditional) expected flow transfer time.*

9.5. PERFORMANCE ANALYSIS

The analytical evaluation of the WLAN performance is split into two stages. STAGE I concentrates on the packet level dynamics at the MAC layer, generalising the analysis first presented in [26] (and subsequently improved by [225]), by incorporating the possibility of capture at the physical layer. The outcome of STAGE I is the aggregate system throughput as a function of the number of persistently active data flows. STAGE II then focuses on the flow level performance using a GPS queueing model, and includes the impact of the dynamics of flow arrivals and departures. At this level, for the analysis of e.g. the flow transfer delay we utilise the aggregate system throughput function as provided by STAGE I. Such a decomposed packet/flow level analysis can be expected to work well if the level-specific dynamics occur at sufficiently distinct time scales.

9.5.1. THROUGHPUT ANALYSIS FOR PERSISTENT FLOWS

The STAGE I analysis builds upon the approach presented in [26, 225] and generalises the considered model with the incorporation of packet capture in case of concurrent transfers. In a scenario with n persistent data flows, the MAC layer operations of a single tagged data flow are modelled by a Markov chain, while the impact of the other $n-1$ flows is incorporated by means of the packet error probability $\mathbf{P}_e$. In turn, from the equilibrium distribution, which is expressed in terms of the Markov chain's characterising parameter $\mathbf{P}_e$, an expression for the packet transfer probability $\mathbf{P}_t^\star$ of an individual flow can be derived, requiring the numerical determination of a unique fixed point. Subsequently, the equilibrium distribution is utilised to derive a closed-form expression for the expected aggregate system throughput.

A key approximation that is made in the analysis is the independence of the different flows' transfer events, which implies that the packet error probability is independent of the number of transfer reattempts the tagged data flow required thus far (cf. [26]). In practice, when a tagged flow's data packet collides irrepairably, not only the tagged flow's contention window size is doubled, but typically also that of the interfering data flow, which in turn decreases the probability that the next packet transfer attempt fails as well. Since under capture model CM-2 a colliding packet still has a chance of survival, the independence assumption is expected to have slightly less impact than under the stricter capture model CM-1, depending on the capture threshold z_0.

The evaluation approach is worked out in more detail below.

EMBEDDED MARKOV CHAIN ANALYSIS

Consider a single tagged data flow contending for the WLAN's shared medium with $n-1$ other flows. Denote with $b(t)$ the stochastic process representing the tagged flow's backoff counter, and with $r(t)$ the stochastic process counting the number of transfer reattempts for the packet at the head of the tagged flow's queue that currently awaits a successful transfer. The embedded jump chain following the state transition of the two-dimensional stochastic process $(r(t), b(t))_{t\geq 0}$ is modelled by an irreducible discrete-time Markov chain $(r(k), b(k))_{k\in\mathbb{N}_0}$, with states denoted (r, b), see Figure 9.4. Observe that in each state at the left of the diagram, i.e. with $b = 0$, the tagged flow (re)attempts a packet transfer, while in all states to the right of such a 'transfer state', the tagged flow is decrementing its backoff counter. The contention window size cw_r, as specified in (9.1), sets the upper bound for the randomly sampled initial backoff value. The state space $\mathbb{S}$ of the Markov chain is

$$\mathbb{S} \equiv \{(r, b) \in \mathbb{N}_0 \times \mathbb{N}_0 : 0 \le b \le cw_r - 1 \text{ and } 0 \le r \le r_{\max}\}.$$

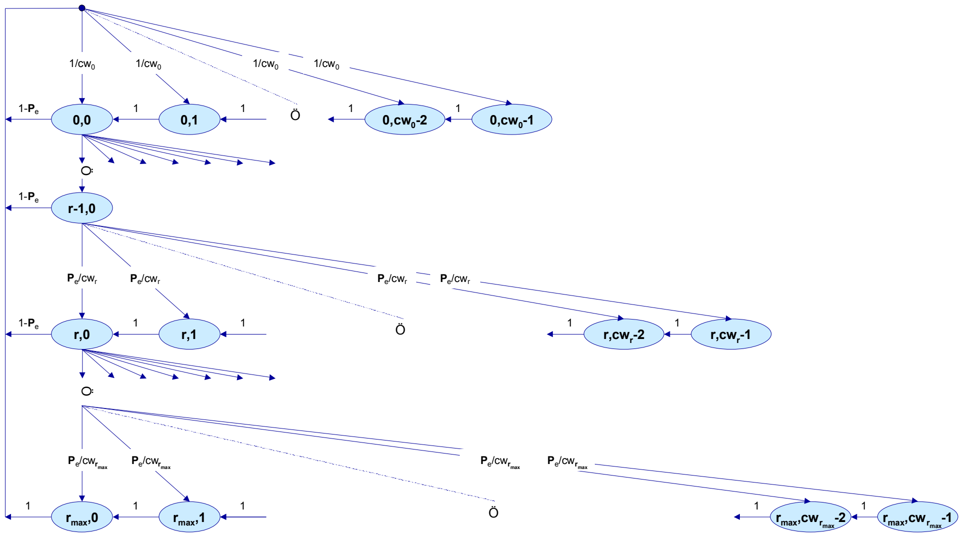


Figure 9.4 The embedded Markov chain of the two-dimensional $(r(t), b(t))_{t\ge 0}$ semi-Markov process, which describes the MAC layer evolution of a single persistent data flow.

The influence of the other $n - 1$ data flows sharing the wireless medium is incorporated in the Markov chain by means of the packet error probability $\mathbf{P}_e$, i.e. the probability that a packet transfer collides irrecoverably with one or more other simultaneous packet transfers. It is stressed that a temporary freeze of the backoff counter due to the sensed activity of another data flow, affects only the *time* between subsequent decrements of $b(t)$, *not* the evolution of the embedded jump chain considered here. This effect is included in the throughput analysis below.

As indicated in Figure 9.4, the Markov chain's one-step transition probabilities are then given by

$$\begin{cases} \Pr\{r,b\,|\,r,b+1\} & = & 1 & \text{for } 0 \le r \le r_{\max}, & 0 \le b \le cw_r - 2, \\ \Pr\{0,b\,|\,r,0\} & = & (1-\mathbf{P}_e)/cw_0 & \text{for } 0 \le r \le r_{\max}-1, & 0 \le b \le cw_0 - 1, \\ \Pr\{r,b\,|\,r-1,0\} & = & \mathbf{P}_e/cw_0 & \text{for } 1 \le r \le r_{\max}, & 0 \le b \le cw_r - 1, \\ \Pr\{0,b\,|\,r_{\max},0\} & = & 1/cw_0 & \text{for} & 0 \le b \le cw_0 - 1, \end{cases}$$

where $\Pr\{r_1,b_1\,|\,r_0,b_0\}$ is short notation for $\Pr\{r(k+1)=r_1, b(k+1)=b_1|r(k)=r_0, b(k)=b_0\}$. The four different event types correspond with a decrement of the backoff counter; a successful packet transfer; an erroneous packet transfer; and a reset of the retransmission counter after the $r_{\max}^{\text{th}}$ MAC layer packet transfer reattempt (regardless of whether the transfer is successful or not), respectively. The last-mentioned event type is where [26] and [225] differ: while in [26] the considered station continues to attempt the packet transfer until it is successful, in [225] the station gives up after $r_{\max}$ reattempts, as is the case in our model. The Markov chain evolution following the second and fourth event type is due to the persistency of the considered data flow. For the MAC layer throughput analysis presented in this section, it is irrelevant whether a packet that suffers from $r_{\max}+1$ unsuccessful transfer attempts is discarded or scheduled for retransmission by higher-layer protocols, due to the assumed *persistent* nature of the considered data flows.

Since the discrete-time Markov chain is irreducible and has a finite state space, a unique equilibrium distribution $(\pi(r,b), (r,b) \in \mathbb{S})$ exists, which can be determined from the global balance equations. Starting e.g. from a state $(r, cw_r - 1)$, the equilibrium probabilities of all states (r,b), $0 \le b \le cw_r - 1$, are readily expressed in terms of $\pi(r-1,0)$ (or $\pi(r_{\max},0)$ in the case of $r=0$). Straightforward recursive manipulations yield (cf. [225])

$$\pi(r,b) = \frac{cw_r - b}{cw_r}\mathbf{P}_e^r \pi(0,0), \tag{9.3}$$

for $0 \leq r \leq r_{\max}$ and $0 \leq b \leq cw_r - 1$, while the normalisation condition for the equilibrium distribution is imposed to determine $\pi(0,0)$:

$$\begin{aligned} 1 &= \sum_{r=0}^{r_{\max}} \sum_{b=0}^{cw_r-1} \pi(r,b) \Leftrightarrow \pi(0,0) \left(\frac{1}{2} \sum_{r=0}^{r_{\max}} (cw_r + 1) \mathbf{P}_e^r \right) = 1 \\ \Leftrightarrow &\quad \pi(0,0) = \frac{2(1-\mathbf{P}_e)}{\left(1-\mathbf{P}_e^{r_{\max}+1}\right) + (1-\mathbf{P}_e) \sum\limits_{r=0}^{r_{\max}} cw_r \mathbf{P}_e^r}. \end{aligned} \tag{9.4}$$

Given the cw_r as specified in (9.1), expression (9.4) can be written more explicitly using

$$\sum_{r=0}^{r_{\max}} cw_r \mathbf{P}_e^r = \begin{cases} (cw_{\min}+1) \left(\dfrac{1-(2\mathbf{P}_e)^{r_{\max}+1}}{1-2\mathbf{P}_e} \right) & \text{if } r_{\max} \leq r^\star, \\ (cw_{\min}+1) \left(\dfrac{1-(2\mathbf{P}_e)^{r^\star+1}}{1-2\mathbf{P}_e} + 2^{r^\star} \dfrac{\mathbf{P}_e^{r^\star+1} - \mathbf{P}_e^{r_{\max}+1}}{1-\mathbf{P}_e} \right) & \text{if } r_{\max} > r^\star, \end{cases}$$

The equilibrium distribution is then completely specified by (9.3) and (9.4) as a function of the (still unknown) packet error probability $\mathbf{P}_e$.

The next step is to express $\mathbf{P}_e$ in terms of the equilibrium distribution derived for a tagged data flow. Firstly, we derive the equilibrium probability $\mathbf{P}_t^\star$ that a specific flow attempts a data packet transfer at a randomly selected event, given by

$$\mathbf{P}_t^\star = \sum_{r=0}^{r_{\max}} \pi(r,0) = \pi(0,0) \sum_{r=0}^{r_{\max}} \mathbf{P}_e^r = \frac{1-\mathbf{P}_e^{r_{\max}+1}}{1-\mathbf{P}_e} \pi(0,0). \tag{9.5}$$

In a system with n data flows, the probability that a tagged data packet is erroneous can be determined by conditioning on the number of simultaneous packet transfers:

$$\mathbf{P}_e = \sum_{k=1}^{n} \mathcal{B}(n-1, \mathbf{P}_t^\star, k-1) \, (1 - \mathbf{P}_s^\star(k)), \tag{9.6}$$

where $\mathcal{B}\left(n-1, \mathbf{P}_t^\star, k-1\right)$ denotes the binomial probability that $k-1$ out of $n-1$ *other* stations attempt a packet transfer concurrently. Expression (9.6) utilises the assumed independence of the different flows' packet transfer attempts. The $\mathbf{P}_s^\star\left(k\right)$ are specified in Section 9.4 and depend on the applied capture model. Note that unlike in [96, 97], the effects of capture are incorporated in the dynamics of the Markov chain, and hence influence the equilibrium distribution and, in particular, the packet transfer probability $\mathbf{P}_t^\star$. Observe that if we substitute (9.4) in (9.5), $\mathbf{P}_t^\star$ is expressed in terms of $\mathbf{P}_e$, while $\mathbf{P}_e$ in turn is expressed as a function of $\mathbf{P}_t^\star$ in (9.6).

Proposition 9.1 *A unique tuple $\left(P_t^\star, P_e\right)$ exists which satisfies expressions (9.4), (9.5) and (9.6).*

Proof Expressions (9.5) and (9.6), along with $\pi\left(0,0\right)$ as specified in (9.4), explicitly define functions f and g given by

$$\mathbf{P}_t^\star = f\left(\mathbf{P}_e\right) = 2\left(1+\frac{\sum\limits_{r=0}^{r_{\max}} cw_r \mathbf{P}_e^r}{\sum\limits_{r=0}^{r_{\max}} \mathbf{P}_e^r}\right)^{-1},$$

and

$$\mathbf{P}_e = g\left(\mathbf{P}_t^\star\right) = \sum_{k=1}^{n} \binom{n-1}{k-1} \left(\mathbf{P}_t^\star\right)^{k-1} \left(1-\mathbf{P}_t^\star\right)^{n-k} \left(1-\mathbf{P}_s^\star\left(k\right)\right).$$

In overview, we will prove Proposition 9.1 by showing that f is a non-increasing function from $[0,1]$ to $\left[f\left(1\right), f\left(0\right)\right]$ with $f\left(0\right) = 2/\left(1+cw_0\right) \geq f\left(1\right) = 2\left(r_{\max}+1\right) / \left(\sum_{r=0}^{r_{\max}}\left(1+cw_r\right)\right) > 0$, while g is a non-decreasing function from $[0,1]$ to $\left[0, 1-\mathbf{P}_s^\star\left(n\right)\right]$ with $0 \leq 1-\mathbf{P}_s^\star\left(n\right)$, which by Brouwer's fixed point theorem (e.g. [29]) implies the existence of a unique fixed point.

First we prove that f is non-increasing. Defining

$$\psi\left(\mathbf{P}_e\right) \equiv \frac{\sum\limits_{r=0}^{r_{\max}} cw_r \mathbf{P}_e^r}{\sum\limits_{r=0}^{r_{\max}} \mathbf{P}_e^r},$$

$\mathbf{P}_t^\star$ can be written as

$$\mathbf{P}_t^\star = f\left(\mathbf{P}_e\right) = 2\left(1 + \psi\left(\mathbf{P}_e\right)\right)^{-1}.$$

We will prove that f is non-increasing by showing that ψ is non-decreasing. Observe first that

$$\psi\left(0\right) = cw_0 \leq \psi\left(1\right) = \frac{1}{r_{\max}+1}\sum_{r=0}^{r_{\max}} cw_r,$$

as cw_r is non-decreasing in r. For $\psi\left(0\right) = \psi\left(1\right) = \psi^\star$, since cw_r is non-decreasing in r it must be that $cw_r = cw_0$, $r = 1, \cdots, r_{\max}$, and the function ψ must be constant (and hence indeed non-decreasing). Alternatively, for the case of $\psi\left(0\right) < \psi\left(1\right)$, we need to derive that for an arbitrary $\psi^\star \in \left(\psi\left(0\right), \psi\left(1\right)\right)$, the number of times the function ψ crosses $\psi^\star$ is equal to 1. Notice that

$$\frac{\sum\limits_{r=0}^{r_{\max}} cw_r \mathbf{P}_e^r}{\sum\limits_{r=0}^{r_{\max}} \mathbf{P}_e^r} = \psi^\star \Leftrightarrow \sum_{r=0}^{r_{\max}} \left(cw_r - \psi^\star\right)\mathbf{P}_e^r = 0. \tag{9.7}$$

Since cw_r is non-decreasing in r and $cw_0 < \psi^\star < cw_{r_{\max}}$, $\left(cw_r - \psi^\star\right)$ changes sign precisely once as r runs from 0 to $r_{\max}$. Invoking Descartes' sign rule (cf. [211]), the polynomial in (9.7) has no more than a single positive root, and hence ψ crosses $\psi^\star$ no more than once. The continuity of ψ then implies that ψ must cross $\psi^\star$ precisely once. As a consequence, ψ is indeed a non-decreasing function, and hence f is non-increasing in $\mathbf{P}_e$.

Secondly, we prove that g is non-decreasing, Using $0^0 = 1$, observe from (9.6) that $g\left(0\right) = 1 - \mathbf{P}_s^\star\left(1\right) \leq 1 - \mathbf{P}_s^\star\left(n\right) = g\left(1\right)$, and

$$\begin{aligned}
\frac{d}{d\mathbf{P}_t^\star} g\left(\mathbf{P}_t^\star\right) &= \frac{d}{d\mathbf{P}_t^\star}\left\{\sum_{k=1}^{n}\binom{n-1}{k-1}\left(\mathbf{P}_t^\star\right)^{k-1}\left(1-\mathbf{P}_t^\star\right)^{n-k}\left(1-\mathbf{P}_s^\star\left(k\right)\right)\right\} \\
&= \sum_{k=1}^{n}\left(1-\mathbf{P}_s^\star\left(k\right)\right)\binom{n-1}{k-1}\Big\{\left(k-1\right)\left(\mathbf{P}_t^\star\right)^{k-2}\left(1-\mathbf{P}_t^\star\right)^{n-k} + \\
&\quad -\left(n-k\right)\left(\mathbf{P}_t^\star\right)^{k-1}\left(1-\mathbf{P}_t^\star\right)^{n-k-1}\Big\}
\end{aligned}$$

$$= \sum_{k=1}^{n-1} \left(\mathbf{P}_t^\star\right)^{k-1} \left(1 - \mathbf{P}_t^\star\right)^{n-k-1} \frac{(n-1)!}{(n-k-1)!\,(k-1)!} \left\{\mathbf{P}_s^\star(k) - \mathbf{P}_s^\star(k+1)\right\} \geq 0,$$

for $\mathbf{P}_t^\star \in [0,1]$, since $\mathbf{P}_s^\star(k)$ was assumed to be non-increasing in k. Hence g is non-decreasing in $\mathbf{P}_t^\star$. □

THROUGHPUT ANALYSIS

In order to determine the expected aggregate system throughput, we first need to introduce some additional notation and parameters. Let τ denote the IEEE 802.11 time slot duration and $\mathbf{T}_s$ ($\mathbf{T}_c$) denote the expected inter-event time in case of a successful (erroneous) packet transfer which may or may not include the tagged data flow. In the BASIC access mode, these inter-event times are

$$\begin{cases} \mathbf{T}_s^{\text{BASIC}} = \text{PHY} + \text{MAC} + r_{\text{WLAN}}^{-1}\mathbf{E}\left\{\overline{X}_s\right\} + \delta + \text{SIFS} + \text{ACK} + \delta + \text{DIFS}, \\ \\ \mathbf{T}_c^{\text{BASIC}} = \text{PHY} + \text{MAC} + r_{\text{WLAN}}^{-1}\mathbf{E}\left\{\overline{X}_c\right\} + \delta + \text{DIFS}, \end{cases}$$

where PHY and MAC denote the physical header (plus preamble) and MAC header sizes (converted to seconds), $\mathbf{E}\left\{\overline{X}_s\right\}$ is the expected net payload size (in kbits) of the largest packet involved in a 'fruitful' collision, i.e. where the 'strongest' packet is successfully captured, δ is the propagation delay between sender and receiver (in seconds), ACK is the acknowledgement message size (converted to seconds), and $\mathbf{E}\left\{\overline{X}_c\right\}$ is the expected net payload size of the largest packet involved in a fruitless collision (in kbits). In the RTS/CTS access mode we have

$$\begin{cases} \mathbf{T}_s^{\text{RTS/CTS}} = \text{RTS} + \delta + \text{SIFS} + \text{CTS} + \delta + \text{SIFS} + \text{PHY} + \text{MAC} + r_{\text{WLAN}}^{-1}\mathbf{E}\left\{\overline{X}_s\right\} + \\ \qquad\qquad + \delta + \text{SIFS} + \text{ACK} + \delta + \text{DIFS}, \\ \\ \mathbf{T}_c^{\text{RTS/CTS}} = \text{RTS} + \delta + \text{DIFS}. \end{cases}$$

Although in practice $\mathbf{E}\left\{\overline{X}_s\right\}$, $\mathbf{E}\left\{\overline{X}_c\right\}$ as well as the expected packet payload size $\mathbf{E}\left\{X\right\}$ tend to be slightly smaller than the given packet size, due to the contribution of the typically smaller packet containing a non-persistent flow's remainder, we assume that $\mathbf{E}\left\{\overline{X}_s\right\} = \mathbf{E}\left\{\overline{X}_c\right\} = \mathbf{E}\{X\}$ is equal to the given packet size, in line with the assumption made in [225]. The values of DIFS, SIFS, PHY, MAC, δ, RTS, CTS, ACK, τ, r_{WLAN} and the packet size are specified in Section 9.6.

From a *single flow*'s perspective in a system with n persistent data flows, the expected call-average data throughput is equal to the *event rate* × *the fraction of events that correspond with successful packet transfers (for the considered flow)* × *the expected transfer volume in case of a successful packet transfer:*

$$\begin{aligned}\beta(n) &\equiv (\mathbf{E}\{\text{inter-event time}\})^{-1}\,\mathbf{P}_t^\star(1-\mathbf{P}_e)\,\mathbf{E}\{X\}\\ &= \frac{\frac{1}{n}\sum\limits_{k=1}^{n}\mathcal{B}(n,\mathbf{P}_t^\star,k)\,\mathbf{P}_s(k)\,\mathbf{E}\{X\}}{\mathcal{B}(n,\mathbf{P}_t^\star,0)\,\tau+\sum\limits_{k=1}^{n}\mathcal{B}(n,\mathbf{P}_t^\star,k)\left\{\mathbf{P}_s(k)\,\mathbf{T}_s+(1-\mathbf{P}_s(k))\,\mathbf{T}_c\right\}},\end{aligned}$$

(in kbits/s) where the expected inter-event time (the inverse of the event rate) is determined by conditioning on the occurrence of three different event types: *(i)* none of the data flows attempts a packet transfer; *(ii)* some of the data flows attempt a packet transfer and the data packet with the strongest signal is successfully captured by the intended receiver; *(iii)* some of the data flows attempt a packet transfer which all collide irrepairably. Note that the duration of a temporary freeze of the considered flow's backoff counter is incorporated in the expected inter-event time in the denominator, i.e. those times when the considered data flow does not attempt a packet transfer but one or more other data flows do.

To conclude this section, the expected *aggregate* system throughput is equal to $n\beta(n)$. Observe that in the simple case of capture model CM-1, the expression for the aggregate system throughput simplifies to

$$\begin{aligned}n\beta(n) &= \frac{\mathcal{B}(n,\mathbf{P}_t^\star,1)\,\mathbf{E}\{x\}}{\mathcal{B}(n,\mathbf{P}_t^\star,0)\,\tau+\mathcal{B}(n,\mathbf{P}_t^\star,1)\,\mathbf{T}_s+\sum\limits_{k=2}^{n}\mathcal{B}(n,\mathbf{P}_t^\star,k)\,\mathbf{T}_c}\\ &= \frac{n\mathbf{P}_t^\star(1-\mathbf{P}_t^\star)^{n-1}\,\mathbf{E}\{x\}}{\left\{\begin{array}{c}(1-\mathbf{P}_t^\star)^n\,\tau+n\mathbf{P}_t^\star(1-\mathbf{P}_t^\star)^{n-1}\,\mathbf{T}_s+\\ +\left(1-(1-\mathbf{P}_t^\star)^n-n\mathbf{P}_t^\star(1-\mathbf{P}_t^\star)^{n-1}\right)\mathbf{T}_c\end{array}\right\}},\end{aligned}$$

which is identical to the aggregate system throughput expression given in [26, 225].

9.5.2. TRANSFER TIME ANALYSIS FOR NON-PERSISTENT FLOWS

In STAGE II we consider the WLAN from the flow level point of view as a service center serving flows at varying rates depending on the number of stations simultaneously active. In particular, when n stations are active the service rate per flow (station) is $\beta\,(n)$, as derived in the previous section for the situation with n persistently active flows, $n = 1, ..., n_{\max}$. The resulting model is a Generalised Processor Sharing queueing model with state-dependent service rates and a finite number of service positions. Assuming, as in our case, that the time instants at which new flow transmissions start constitute a Poisson process, this GPS model is analytically tractable. In particular, the equilibrium distribution of the number of flows simultaneously in progress is given by

$$\widetilde{\pi}\,(n) = \frac{\rho^n \phi\,(n)}{\sum\limits_{n'=0}^{n_{\max}} \rho^{n'} \phi\,(n')} \quad \text{with} \quad \phi\,(n) \equiv \left(\prod_{n'=1}^{n} \frac{n' \beta\,(n')}{r_{\text{WLAN}}} \right)^{-1},$$

$n = 1, ..., n_{\max}$ (see [54]), where $\rho \equiv \lambda/(\mu r_{\text{WLAN}})$ and $\phi\,(0) \equiv 1$ by convention.

From the equilibrium distribution we can compute the expected number of flows present in the system and, using Little's formula [213], the expected flow transfer time $\mathbf{T}$:

$$\mathbf{T} \equiv \sum_{n=0}^{n_{\max}} \frac{n\,\widetilde{\pi}\,(n)}{\lambda\,(1 - \widetilde{\pi}\,(n_{\max}))},$$

Some additional interesting results for the GPS model have been derived (see [54]). In particular, the conditional expected transfer time $\mathbf{T}\,(x)$ of a flow of given size $x \geq 0$ can be computed explicitly and grows linearly in x:

$$\mathbf{T}\,(x) \equiv \frac{x}{r_{\text{WLAN}}} \sum_{n=0}^{n_{\max}} \frac{n\,\widetilde{\pi}\,(n)}{\rho\,(1 - \widetilde{\pi}\,(n_{\max}))},$$

a result which expresses the fair allocation of WLAN capacity to the served flows.

An important feature of the GPS model is that these performance measures are *insensitive* with respect to the specific form of the flow size distribution, depending on its mean $1/\mu$ only. These attractive properties suggested by our modelling approach are validated by simulation results of the WLAN system in the next section.

9.6. NUMERICAL RESULTS

In order to validate the presented analysis as well as to obtain valuable insights into the WLAN performance, we now present a number of numerical experiments that are obtained by means of analytical evaluations and dynamic simulations. The simulation program features the specified traffic and capture models as well as a detailed representation of the IEEE 802.11 DCF MAC protocol. Numerical results are presented both for the STAGE I throughput analysis for persistent data flows, as well as for the STAGE II transfer time analysis for non-persistent data flows. The MAC and DSSS physical layer parameter settings are outlined in Table 9.1, along with the relevant traffic and capture model parameters. Recall that the physical layer preamble and header are always transmitted at 1 Mbits/s, while all other messages are transferred at the applied channel rate, so that the duration of e.g. an ACK frame takes $192/10^3+112/r_{\text{WLAN}}$ ms. The considered range for z_0 is taken from [96, 97], in order to allow a comparison of the obtained aggregate system throughput. As default parameter settings we use a channel rate of $r_{\text{WLAN}} = 1$ Mbits/s, the associated spreading factor of $S_f = 11$, and a required energy-per-bit to interference-plus-noise density ratio of $z_0 = 15$ dB for the case of capture model CM-2. Sufficient independent replications were carried out to obtain 95% confidence intervals with a relative precision no worse than 5%.

9.6.1. THROUGHPUT RESULTS FOR PERSISTENT FLOWS

Consider the aggregate system throughput (in Mbits/s) as a function of the number of persistent flows. Figure 9.5 (left) shows for the BASIC access mode both the scenario without (CM-1) and with (CM-2) capture, while Figure 9.5 (right) shows the corresponding results for the RTS/CTS access mode (note the difference in the value range on the vertical axes). Recall that for the situation without capture, the analytical model simplifies to that of [225]. In the alternate case with capture, the analytical results from the presented model are compared not only with the simulation results, but also with the results obtained from the analytical approach presented in [96, 97]. At this point we note that in contrast with our analysis, in [96, 97] it is implicitly (and incorrectly [90, 95]) allowed that more than only the strongest signal may survive a

Table 9.1 System, traffic and capture model parameter settings for the WLAN performance analysis, based on the DSSS physical layer.

SYSTEM MODEL			TRAFFIC MODEL		
r_{WLAN}	$\in \{1, 2, 5.5, 11\} \cdot 10^3$	kbits/s	μ^{-1}	120	kbits
S_f	$\in \{11, 11, 8, 8\}$	-	ρ	$\in [0, 1]$	-
PHY	192	bits	$n_{\max}$	100	flows
MAC	272	bits	packet size	12	kbits
RTS	PHY + 160	bits			
CTS	PHY + 112	bits	CAPTURE MODEL		
ACK	PHY + 112	bits	z_0	$\in [6, 24]$	dB
δ	1	μs			
τ	20	μs			
SIFS	10	μs			
DIFS	SIFS $+ 2 \times \tau = 50$	μs			
$cw_{\min}$	31	-			
$cw_{\max}$	1023	-			
$r^\star$	5	-			
$r_{\max}$	BASIC: 3	-			
$r_{\max}$	RTS/CTS: 6	-			

collision, requiring only that the carrier-to-interference ratio requirement is met. For $\Gamma^\star < 1$ this assumption has a positive effect on the aggregate system throughput.

Figure 9.5 reveals that for both access modes and capture models our analytical model appears to capture the behaviour of the WLAN extremely well, as was also observed in [225] for the scenario without capture. In a comparison with the analysis of [96, 97] (HVS model) for the scenario with capture, our analytical model outperforms that of HVS for both access modes, while the HVS model still leads to a good yet somewhat too conservative throughput estimates, despite the above-mentioned assumed possibility of multiple packets surviving a collision.

For the RTS/CTS access mode the aggregate system throughput increases for small numbers of persistent data flows but decreases for larger numbers. The rationale behind this lies in the conflicting effects that as the number of persistent data flows increases, the average idle times between transmission attempts decreases, while the

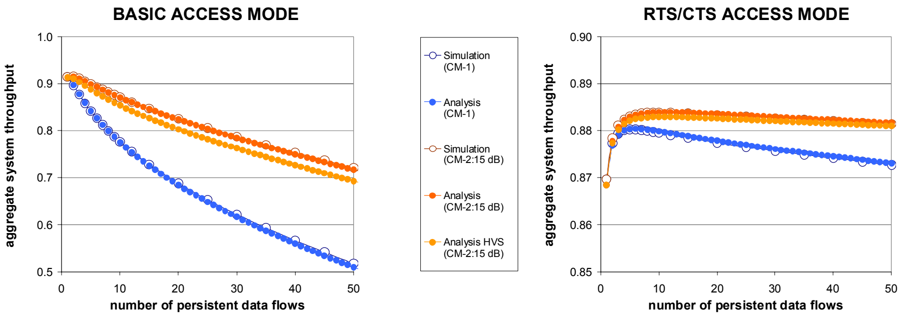


Figure 9.5 STAGE I performance and validation: aggregate system throughput versus the number of persistent data flows for the BASIC (left) and RTS/CTS (right) access modes.

likelihood of collisions increases. Apparently, the former effect dominates for relatively few persistent data flows, while the reduced efficiency caused by the latter effect dominates for a larger number of persistent data flows. In principle, the same trade-off applies to the BASIC access mode, although it is clear from the numerical results that the efficiency loss due to collisions is larger.

Although for a very small number of persistent data flows the BASIC access mode yields a slightly better performance due to its relatively low overhead, the impact of an increasing number of persistent data flows on the aggregate system throughput is much more significant for the BASIC than for the RTS/CTS access mode, due to the inherently different impact of a collision. For a moderate to large number of persistent data flows, the additional overhead of the RTS/CTS access mode clearly pays off in the sense that the aggregate system throughput decreases much less dramatically. Our results suggest the hybrid deployment of the BASIC and RTS/CTS access modes, where the DCF switches between the BASIC to RTS/CTS access modes at a presence of around 10 data flows, in order to optimise the aggregate system throughput. Observe further that the aggregate system throughput is substantially smaller than the assumed channel rate of 1 Mbits/s, which is due to the inefficiency on the MAC layer caused by waiting times, non-data packets and destructive collisions. As we will see below, this relative inefficiency is even worse for higher channel rates, since the DIFS, SIFS and PHY durations are independent of the channel rate.

As can be seen from both charts, the effect of capture on the aggregate system throughput is significant: for the assumed capture threshold of $z_0 = 15$ dB in the case of the BASIC access mode the aggregate system throughput improves up to 40%. For the RTS/CTS access mode the throughput gain from of capture is much smaller, a direct consequence of the relatively small impact of collisions on the aggregate system throughput even in the case without capture (precisely the objective of the RTS/CTS mode).

9.6.2. TRANSFER TIME RESULTS FOR NON-PERSISTENT FLOWS

In STAGE II we evaluate the expected file transfer times on the flow level, corresponding to the more realistic scenario where stations initiate and terminate data flows at random time instances. Consider first the expected flow transfer time (in seconds) versus the data traffic load ρ. For the scenario with capture, Figure 9.6 depicts results for the BASIC (left) and RTS/CTS (right) access modes for different flow size distributions: deterministic, exponential and two hyperexponential distributions with 'balanced means' (see [213]) and coefficient of variation 2 and 4.

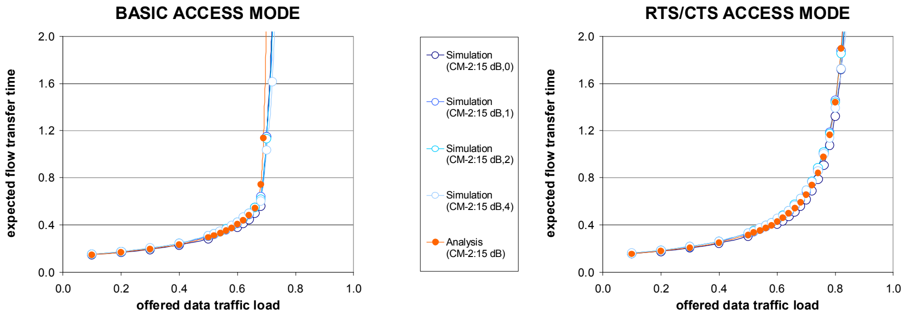


Figure 9.6 STAGE II performance and validation: expected flow transfer time versus the data traffic load for the BASIC (left) and RTS/CTS (right) access modes.

As is clear from the charts, our analytical model approximates the flow level WLAN behaviour excellently for both access modes. The expected flow transfer time increases gradually up to a data traffic load of 0.6 (BASIC access mode) and 0.7 (RTS/CTS access mode), beyond which the expected flow transfer times increase rapidly as the expected number of present data flows approaches the CAC threshold. Note that these values are significantly less than 1 due to the inherent inefficiencies of the CSMA/CA-based multiple access protocol. As the number of present data flows increases, not only does the available throughput per flow decrease due to a fiercer competition for resources, but also the aggregate system throughput itself decreases as is clear from the STAGE I analysis (see Figure 9.5). For the RTS/CTS access mode this performance degradation appears to be less dramatic than for the BASIC access mode, which is directly reflected in a larger 'stability regime'. Strictly speaking, the inclusion of CAC trivially ensures system stability for any data traffic load. With 'stability regime' we refer to the critical data traffic load beyond which a dramatic performance reduction is observed. This critical data traffic load obviously depends on the $\beta(n)$, $n = 1, \cdots, \infty$. In this light we note that in a system with an infinite number of service positions (no CAC), the appropriate stability condition is given by [54]

$$\sum_{n=0}^{\infty} \rho^n \phi(n) < \infty,$$

which for the egalitarian Processor Sharing model (with $\phi(n) \equiv 1$) is readily seen to be equivalent to $\rho < 1$.

Observe that for light to moderate traffic loads the performance of both access modes is similar, while for very light traffic loads the BASIC access mode even performs slightly better, which is line with the STAGE I results. Figure 9.6 further reveals the important result that the expected flow transfer time appears to be insensitive to the specific form of the flow size distribution, which supports the analytically suggested insensitivity property (see Section 9.5.2).

As a second experiment, consider the conditional expected flow transfer time as a function of the flow size for a data traffic load of $\rho = 0.6$, again for the scenario with capture (see Figure 9.7). Aside from the individual flow transfer times observed during the simulation which are scattered against the corresponding flow sizes, the figure also contains the averaged flow transfer times. In order to determine these averages, we have split the range of flow sizes in disjoint segments, and calculated the average flow transfer times for each segment, using the available samples. The latter curves not only support the analytically claimed insensitivity property, but moreover, the simulation results confirm the linearity of the conditional expected flow transfer time in the flow size. Once again, the closeness of the analytical and simulation results indicates that the analytical model provides an excellent approximation for the WLAN behaviour.

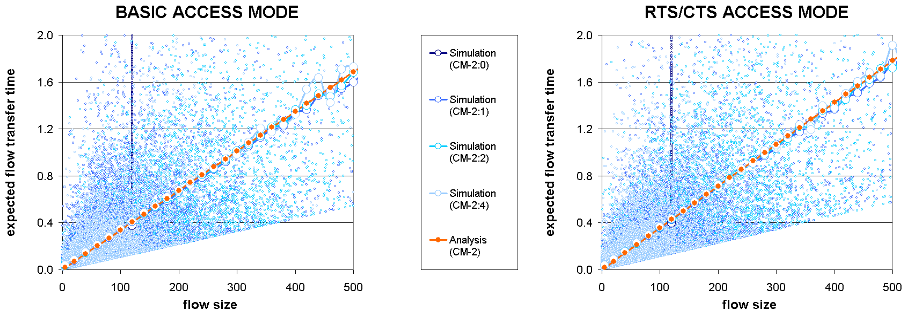


Figure 9.7 STAGE II performance and validation: conditional expected flow transfer times versus the data flow size for the BASIC (left) and RTS/CTS (right) access modes.

As the validation of the analysis has revealed that it evaluates the system behaviour to an excellent precision, despite the number of approximating assumptions made to enable tractability, the remaining experiments are based on the analytical model only. For the scenario with capture, Figures 9.8 and 9.9 depict the expected flow transfer time as a function of the data traffic load ρ and the capture threshold z_0, for different channel rates r_{WLAN} and for the BASIC and RTS/CTS access modes, respectively. Recall that the data traffic load definition incorporates a normalisation with respect to the channel rate.

Observe first the intuitively expected trends that the expected flow transfer time is increasing in both ρ, which determines the degree of competition for resources, and the capture threshold z_0, which determines the likelihood of surviving a collision. In the RTS/CTS access mode, however, the flow level performance is barely affected by z_0. The vertical range of the charts is set sufficiently large to reveal the effects of CAC. Since the CAC threshold is identical in all considered cases, the expected flow transfer times in the worst-case scenario of heavy traffic and poor capture potential, are much lower for higher channel rates. However, as revealed most clearly for the RTS/CTS access mode, the 'critical' data traffic load beyond which the flow level performance worsens significantly, is lower for higher channel rates. This is a due to the fact that the time durations of the DIFS, SIFS and PHY are independent of the channel rate (cf. [110, 111]) and hence place a larger relative claim on the resources if the channel rates are higher. As a consequence, a range of data traffic loads exists where the flow level performance *degrades* for higher channel rates, while in e.g. a standard

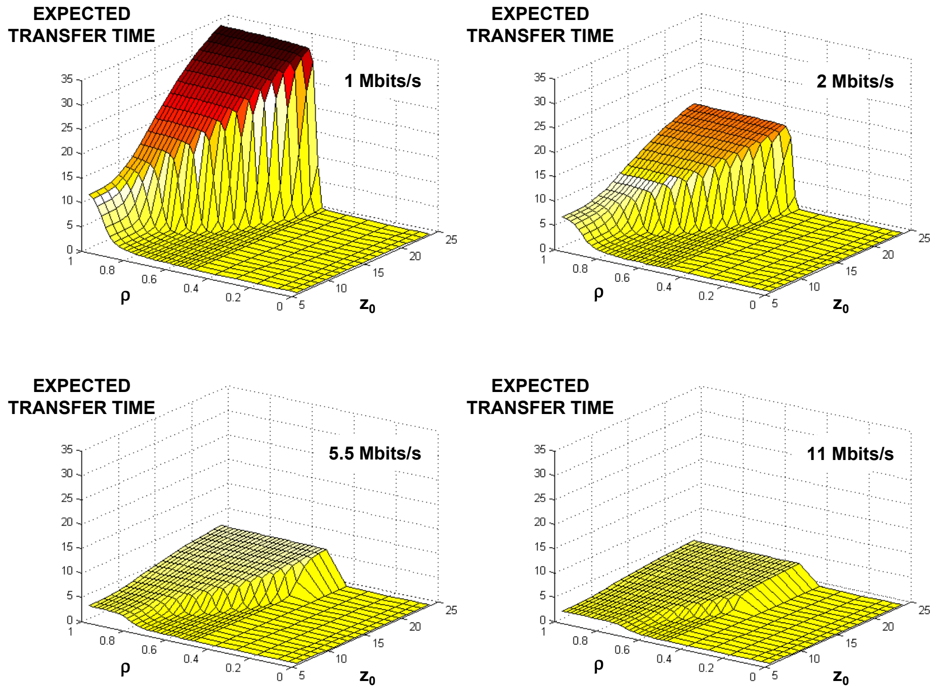


Figure 9.8 STAGE II performance (BASIC access mode): the expected flow transfer time as a function of the data traffic load ρ, the capture threshold z_0, and the channel rate r_{WLAN}.

$M/G/1/PS$ queueing model, the expected flow transfer time is *inversely* proportional to the channel rate. Finally, as observed before, the 'stability regime' of the RTS/CTS access mode is generally larger than that of the BASIC access mode, except for low z_0, i.e. when the resource waste due to collisions and hence the benefit of a four-way handshake is negligible.

9.7. CONCLUDING REMARKS

This chapter has presented an integrated packet/flow level modelling approach for performance evaluation of a IEEE 802.11 WLAN with non-persistent traffic sources. At the packet level, different physical and MAC layer system aspects are integrated into a single model, while still allowing explicit analytical evaluation. The flow level model is based on the observation that a WLAN system behaves approximately as a

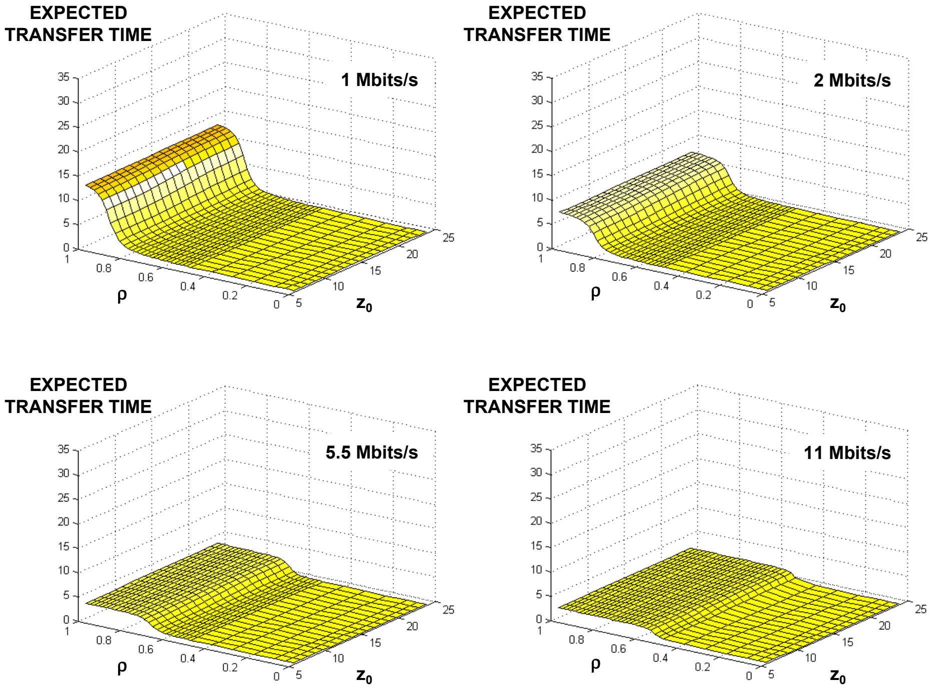


Figure 9.9 STAGE II performance (RTS/CTS access mode): the expected flow transfer time as a function of the data traffic load ρ, the capture threshold z_0, and the channel rate r_{WLAN}.

queueing system with a Generalised Processor Sharing service discipline. Exploiting known performance results for GPS queues we have derived an analytical approximation for (conditional) expected flow transfer times. The accuracy of the approximation has been investigated by comparison of the analytical results with results obtained by simulation. The principal conclusion from the numerical experiments is that the approximation yields very accurate results for all considered scenarios. The approximation very well reflects the severe increase of the expected flow transfer times when the offered traffic load approaches the maximum system throughput. Further, the numerical results show that the positive effect of packet capture on WLAN system throughput and flow transfer times is considerable, yet often ignored in WLAN performance studies. Our modelling approach also provides interesting general insights in the performance characteristics of WLANs. In particular, known results for the

GPS model imply that expected flow transfer times are insensitive to the flow size distribution, which is confirmed by the simulation results.

A numerical comparison of the BASIC and RTS/CTS access modes of the IEEE 802.11 Distributed Coordination Function backs the intuitive expectations and design objectives. Although for a relatively light data traffic load, the RTS/CTS access mode performs slightly worse than the BASIC access mode, due to the resources 'wasted' in the four- rather than two-way handshake, for moderate to heavy traffic loads, this additional overhead pays off in terms of a strongly reduced impact of packet collisions and, correspondingly, a significantly better performance. Consequently, the RTS/CTS access mode can satisfactorily handle a larger regime of data traffic loads. Numerical results further showed that the impact of packet capture is largest under the BASIC access scheme, which is a direct consequence of the inherently large impact of collisions; in other words, there is more room for improvement.

Besides providing valuable input for capacity planning, such numerical experiments are also readily exploited to optimise controllable system parameters, such as the packet size threshold, beyond which the system automatically switches from BASIC to RTS/CTS access mode. Another interesting application of our approximation that deserves more attention, is the optimisation of Call Admission Control. In particular, the analytical model enables swift determination of suitable flow admission thresholds for given requirements on expected flow transfer times, expected throughputs and flow blocking probabilities. Further topics for continued research are the inclusion of TCP flow control, the investigation of the automatic rate fallback mechanism and the MAC layer QOS differentiation mechanisms that are specified by task group TG-E (cf. [3, 112, 162]).

CHAPTER 10

EPILOGUE

AS a concluding chapter of this monograph, the epilogue briefly reviews the key results that have been obtained and provides an outlook for further investigations on capacity allocation in wireless communication networks.

10.1. REVIEW

This monograph has concentrated on capacity allocation in cellular and Wireless Local Area Networks, primarily with a network operator's perspective. In the introductory chapter, a reference model has been proposed for the extensive suite of capacity allocation mechanisms that can be applied at different time scales, in order to influence the inherent trade-offs between investment costs, network capacity and service quality. The subsequent chapters presented a number of comprehensive studies with the objective to understand the joint impact of the different control mechanisms on the network operations and service provisioning, as well as the influence of the largely uncontrollable traffic and mobility characteristics on the system- and service-level performance.

In Chapters 2 through 5 we have developed and evaluated generic yet tractable models for the performance analysis of second-generation integrated services GSM/HSCSD/GPRS networks. A number of relevant Grade and (conditional) Quality Of Service measures have been derived for speech, video, high- and low-priority data services, which enables operators to adequately dimension and control their networks (Chapters 2 and 3). In a sensitivity analysis regarding the impact of the data call size distribution that is motivated by the generally acknowledged observation that e.g. WWW page sizes are heavy tailed, we have presented and analytically supported the counterintuitive phenomenon that in an integrated services model, a greater data call size variability leads to a better QOS (Chapter 4). Another key result is that across a diverse range of topical integrated services models, the newly introduced and readily derived expected instantaneous throughput is demonstrated to be the only one among a variety of more commonly applied throughput measures, that excellently approximates the generally hard-to-determine call-average throughput, which is undeniably the most relevant throughput measure from the users' perspective (Chapter 5).

Chapters 6 through 8 have concentrated on distinct performance issues in third-generation UMTS(/HSDPA) networks, which are inherently more complex to analyse due to the CDMA-based radio interface. Chapter 6 has presented a tractable semi-analytical model for the evaluation of data transfers over Downlink Shared CHannels in multi-cellular data-only UMTS networks, which was utilised to assess the impact of different interference aspects on the experienced data QOS. Chapter 7 has concentrated on a UMTS/HSDPA network, integrating prioritised speech traffic and delay-tolerant data traffic. For this setting, we have developed a method to evaluate the performance of either resource- or QOS-fair adaptive scheduling schemes, combining analytical techniques with Monte Carlo simulations. Among the principal conclusions, it was demonstrated that adaptive scheduling of data transfers allows the enhancement of both speech and data QOS, while the considered fairness alternatives appeared to differ significantly in the delivered QOS and the induced spatial heterogeneity of data traffic. Chapter 8 presented an analytical evaluation method to assess the impact of terminal mobility on UMTS network planning. Focussing on bottom-line performance measures, among the primary insights of the presented numerical experiments terminal mobility was identified as a key property that should not be neglected in the network planning process. Furthermore the deployment of a Radio Resource Reservation scheme in support of mobility-induced handovers appears most effective in reducing premature call termination as well as network investment costs.

Finally, in Chapter 9 we presented an integrated packet/flow-level model for the evaluation of WLAN performance. Closed-form expressions have been derived for both the packet-level aggregate system throughput and the flow-level data transfer times. As the validation of the tractable model by means of extensive simulations turned out to be very successful, the derived expressions have been applied to generate both qualitative and quantitative insight into the impact of different system and traffic model parameters on the WLAN performance.

10.2. OUTLOOK

The unpredictability of the traffic, terminal mobility and propagation characteristics, the on-going development of new services with distinct QOS requirements, the complex interactions of the capacity allocation mechanisms and the continuous technological innovations ensure that numerous sensible investigations on capacity allocation in wireless communication networks are possible.

Concerning the pursued *evaluation methods*, we suspect that if the current cellular network evolution from the second- to the third-generation is representative for further technological advancements with regard to the complexity of analysis, the fruitful applicability of purely *analytical* methods will be limited to indicating general trends and providing qualitative insights, and possibly to more conclusive analyses of suitably isolated problems. A transition of evaluation strategies from a healthy blend of analysis and *simulations* that FD/TDMA-based second-generation networks allowed, towards a predominantly simulation-based approach in the investigation of CDMA-based third-generation networks, is apparent both in the literature and, to some extent, also in this monograph. An important underlying rationale is that radio access technologies have advanced to a level where resources can be efficiently allocated to calls in accordance with their specific circumstances, e.g. the associated terminals' locations and the local traffic load. Moreover, the inherent interactions between the different operational layers and time scales severely complicate the performance analysis and optimisation, which is particularly true for wireless communication networks. As demonstrated in the UMTS-specific investigations in this monograph, which arguably balance on the edge of what is analytically feasible, the inclusion of such aspects invariably encumbers worthwhile analyses. Still, even if conclusive investigations are likely to require a simulation-based approach, an analytical perspective as well as a thorough pre-evaluation of isolated subproblems can provide indispensable assistance in the design of well-targeted experiments and the adequate assessment of the obtained results.

With regard to the relevant *research topics* in the field of capacity allocation in wireless communication networks that deserve attention in the near future, we confine ourselves here to more general suggestions that are not direct extensions of the presented work, as those propositions have been included where applicable in the various 'concluding remarks' sections. As the wireless service offerings are slowly becoming more diverse and non-speech traffic is gradually gaining a significant market share, the development of *traffic* and *mobility models* based on actual measurement traces is now a viable research suggestion, where the correlation between terminal mobility and the type of service is of particular interest. Further, the multitude of studies on capacity allocation in wireless networks concentrates on the radio interface operations alone, ignoring the potential impact of *higher-layer protocols* such as end-to-end TCP flow control and adaptive audio or video codecs. As such protocols may significantly influence both the resource utilisation and the experienced service quality, depending on the considered scenario, it is suggested that these effects are further investigated. A closely associated suggestion is to gain comprehension of the *human perception* of the delivered QOS, which is typically non-linear in the QOS measures that can be derived at the network level. Such understanding helps assess what QOS enhancements due to an optimised deployment of capacity allocation mechanisms are truly appreciated by the customer. A final area that deserves attention is the coordination of capacity allocation in *integrated networks*, e.g. WLANs as hot spot 'islands in a sea' of wide area cellular GSM/HSCSD/GPRS or UMTS coverage. An important new mechanism in such integrated networks is the appropriate assignment of calls to either network type upon call establishment or during a call by means of (seamless) directed vertical handovers (load balancing). A proper assignment policy should take into account e.g. the traffic characteristics and QOS requirements of the requested call, the terminal's mobility and the current loading in each of the networks.

REFERENCES

[1] 3GPP TS 25.211, *"Physical channels and mapping of transport channels onto physical channels (FDD)"*, v3.5.0, Release 99, 2000.

[2] 3GPP TS 25.848, *"Physical layer aspects of UTRA High Speed Downlink Packet Access"*, v4.0.0, Release 4, 2001.

[3] I. Aad and C. Castelluccia, "Differentiation mechanisms for IEEE 802.11", *Proceedings of IEEE INFOCOM '01*, Anchorage, USA, 2001.

[4] W. Ajib and P. Godlewski, "Service disciplines performance for WWW traffic in GPRS system", *Proceedings of 3G Mobile communication technologies*, London, UK, pp. 431-435, 2000.

[5] E. Altman, D. Artiges and K. Traorem, "On the integration of best-effort and guaranteed performance services", *Research report 3222,* INRIA, France, 1997.

[6] A.T. Andersen, S. Blaabjerg, G. Fodor and M. Telek, "A partially blocking-queueing system with CBR/VBR and ABR/UBR arrival streams", *Telecommunication systems*, vol. 19, no. 1, pp. 75-99, 2002.

[7] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar and P. Whiting, "Data rate scheduling algorithms and capacity estimates for the CDMA forward link", *Technical report BL0112120-990922-32TM,* Bell Labs, Lucent Technologies, USA, 1999.

[8] Å. Arvidsson and P. Karlsson, "On traffic models for TCP/IP", *Proceedings of ITC 16*, Edinburgh, Scotland, pp. 457-466, 1999.

[9] B. Avi-Itzhak and S. Halfin, "Expected response times in a non-symmetric time sharing queue with a limited number of service positions", *Proceedings of ITC 12*, pp. 5.4B.2.1-7, 1988.

[10] H. Balakrishnan, V. Padmanabhan, S. Seshan and R.H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links", *IEEE/ACM Transactions on networking,* vol. 5, no. 6, pp. 756-769, 1997.

[11] N. Bambos, S.C. Chen and D. Mitra, "Channel probing for distributed access control in wireless communication networks", *Proceedings of IEEE GLOBECOM '95*, Singapore, 1995.

[12] N.D. Bambos, S.C. Chen and G.J. Pottie, "Radio link admission algorithms for wireless networks with power control and active link quality protection", *Proceedings of INFOCOM '95,* Boston, USA, vol. 1, pp. 97-104, 1995.

[13] A. Bedekar, S.C. Borst, K. Ramanan, P.A. Whiting and E.M. Yeh, "Downlink scheduling in CDMA data networks", *Proceedings of GLOBECOM '99,* Rio de Janeiro, Brazil, pp. 2653-2657, 1999.

[14] L. Begain, "Scalable multimedia services in GSM-based networks: an analytical approach", *Proceedings of the ITC specialist seminar on Mobile systems and mobility*, Lillehammer, Norway, pp. 73-83, 2000.

[15] J.S. Belrose, "On the birth of wireless telephony", available at `http://www.telecommunications.ca/Wireless_Telephony_-_2.pdf`.

[16] S. Ben Fredj, T. Bonald, A. Proutiere, G. Régnié and J.W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level", *Proceedings of SIGCOMM '01*, San Diego, USA, 2001.

[17] N. Benameur, S. Ben Fredj, F. Delcoigne, S. Oueslati-Boulahia and J.W. Roberts, "Integrated admission control for streaming and elastic traffic", *Proceedings of the 2nd International workshop on Quality of future Internet services,* Coimbra, Portugal, 2001.

[18] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger, "Long-range dependence in variable-bit-rate video traffic", *IEEE Transactions on communications*, vol. 43, no. 2-4, pp. 1566-1579, 1995.

[19] J.L. van den Berg, "Sojourn times in feedback and processor-sharing queues", Ph.D. thesis, Rijksuniversiteit Utrecht, The Netherlands, 1990.

[20] J.L. van den Berg and O.J. Boxma, "The $M/G/1$ queue with processor sharing and its relation to a feedback queue", *Queueing systems,* vol. 9, pp. 365-401, 1991.

[21] J.L. van den Berg, R.D. van der Mei, B.M.M. Gijsen, M.J. Pikaart and R. Vranken, "Processing times for transaction servers with quality of service differentiation", *Proceedings of the 11th GI/ITG conference on Measuring, modelling and evaluation of computer and communications systems*, Aachen, Germany, pp. 73-83, 2001.

[22] F. Berggren and S.-L. Kim, "Energy-efficient downlink power control and scheduling for CDMA non-real time data", *Proceedings of MMT '00,* Duck Key, USA,

2000.

[23] R. De Bernardi, D. Imbeni, L. Vignali and M. Karlsson, "Load control strategies for mixed services in WCDMA", *Proceedings of VTC '00*, pp. 825-829, 2000.

[24] U. Bernhard, H. Pampel, J. Mueckenheim and P. Gunreben, "Evaluation of W-CDMA network performance and impact of soft handover using dynamic network simulations", *Proceedings of 3G Mobile communication technologies '00*, London, UK, pp. 347-351, 2000.

[25] C. Bettstetter, H.-J. Vögel and Jörg Eberspächer, "GSM phase 2+ General Packet Radio Service: architecture, protocols, and air interface", *IEEE Communications surveys*, `http://www.comsoc.org/pubs/surveys`, vol. 2 no. 3, 1999.

[26] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function" , *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535-547, 2000.

[27] G. Bianchi, A. Capone, L. Fratta and L. Musumeci, "Packet data service over GSM networks with dynamic stealing of voice channels", *Proceedings of GLOBECOM '95*, pp. 1152-1156, 1995.

[28] G. Bianchi, A. Capone, L. Fratta and L. Musumeci, "Voice and packet data integration over GSM networks", *Proceedings of ICCC '95*, Seoul, South Korea, pp. 297-302, 1995.

[29] K.G. Binmore, *"Mathematical analysis: a straightforward approach"*, Cambridge University Press, Cambridge, UK, 1982.

[30] T. Bonald and L. Massoulié, "Impact of fairness on Internet performance", *Proceedings of SIGMETRICS '01*, Cambridge, USA, 2001.

[31] T. Bonald and J.W. Roberts, "Performance of bandwidth sharing mechanisms for service differentiation in the Internet", *Proceedings of the ITC specialist seminar on IP traffic measurement, modelling and management*, Monterey, USA, pp. 22.1-22.10, 2000.

[32] N.K. Boots, *"Rare event simulation in models with heavy-tailed random variables"*, Ph.D. thesis, Vrije Universiteit Amsterdam, The Netherlands, 2002.

[33] S.C. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks", *Proceedings of INFOCOM '03,* San Francisco, USA, 2003.

[34] S.C. Borst, M.J.G. van Uitert and M.R.H. Mandjes, "GPS queues with heterogeneous traffic classes", *Proceedings of INFOCOM '02,* New York, USA, 2002.

[35] S.C. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimisation", *Proceedings of INFOCOM '01,* Anchorage, USA, 2001.

[36] R.J. Boucherie and M. Mandjes, “Estimation of performance measures for product form cellular mobile communications networks”, *Telecommunication systems*, vol. 10, no. 3, pp. 321-354, 1998.

[37] G. Brasche and B. Walke, “Concepts, services, and protocols of the new GSM phase 2+ General Packet Radio Service”, *IEEE Communications magazine*, vol. 35, no. 8, pp. 94-104, 1997.

[38] G. Brussaard, Technische Universiteit Eindhoven, The Netherlands, *private communications*, 2003.

[39] J. Cai and D.J. Goodman, “General Packet Radio Service in GSM”, *IEEE Communications magazine*, vol. 35, no. 10, pp. 122-131, 1997.

[40] F. Cali, M. Conti and E. Gregori, “IEEE 802.11 wireless LAN: capacity analysis and protocol enhancement”, *Proceedings IEEE INFOCOM ’98*, San Francisco, USA, 1998.

[41] D. Calin, S. Malik and D. Zeghlache, “Traffic scheduling and fairness for GPRS air interface”, *Proceedings of VTC ’99*, Amsterdam, The Netherlands, pp. 834-838, 1999.

[42] D. Calin and D. Zeghlache, “Performance analysis of High Speed Circuit Switched Data (HSCSD) over GSM”, *Proceedings of ICC ’98*, pp. 1586-1590, 1998.

[43] D. Calin and D. Zeghlache, “High Speed Circuit Switched Data over GSM: potential traffic policies”, *Proceedings of VTC ’98*, pp. 1274-1278, 1998.

[44] D. Calin and D. Zeghlache, “Priority queueing analysis for voice-data integration in wireless PCS”, *Proceedings of MMT ’98*, pp. 273-286, 1998.

[45] Jonathan P. Castro, *“The UMTS network and radio access technology: air interface for future mobile systems”*, John Wiley & Sons, Chichester, England, 2001.

[46] C.-J. Chang, B.-W. Chen, T.-Y. Liu and F.-C. Ren, “Fuzzy/neural congestion control for integrated voice and data DS-CDMA/FRMA cellular networks”, *IEEE Journal on selected areas in communications*, vol. 18, no. 2, pp. 283-293, 2000.

[47] K. Chawla, P.F. Driessen and X. Qiu, “Transmission of streaming data over an EGPRS wireless network”, *Proceedings of VTC ’00*, Tokyo, Japan, pp. 118-122, 2000.

[48] H.S. Chhaya and S. Gupta, “Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocol”, *Wireless networks*, vol. 3, no. 3, pp. 217-234, 1997.

[49] E. Chlebus, “Empirical validation of call holding time distribution in cellular communication systems”, *Proceedings of ITC 15*, Washington DC, USA, pp.

1179-1188, 1997.

[50] Y.M. Chuang, T.Y. Lee and Y.B. Lin, “Trading CDPD availability and voice blocking probability in cellular networks”, *IEEE Network magazine*, pp. 48-54, March/April 1998.

[51] E.A. Coddington and N. Levinson, *“Theory of ordinary differential equations”*, McGraw-Hill, New York, USA, 1955.

[52] E.G. Coffman, R.R. Muntz and H. Trotter, “Waiting time distributions for processor-sharing systems”, *Journal of the Association for Computing Machinery*, vol. 17, pp. 123-130, 1970.

[53] J.W. Cohen, “Some results on regular variation in queueing and fluctuation theory”, *Journal of applied probability*, vol. 10, pp. 343-353, 1973.

[54] J.W. Cohen, “The multiple phase service network with generalized processor sharing”, *Acta informatica*, vol.12, pp. 245-284, 1979.

[55] C. Cordier and A. de Hoz, “Dimensioning rules for CDMA systems”, *Proceedings of PIMRC ’99*, Osaka, Japan, pp. 940-945, 1999.

[56] M. Crovella and A. Bestavros, “Self-similarity in World Wide Web traffic: evidence and possible causes”, *IEEE/ACM Transactions on networking*, vol. 5, no. 6, pp. 835-846, 1997.

[57] M.E. Crovella and L. Lipsky, “Simulations with heavy-tailed workloads”, Chapter 3 of *“Self-similar network traffic and performance evaluation”* (eds.: K. Park and W. Willinger), John Wiley & Sons, Chichester, England, 2000.

[58] C. Cunha, A. Bestavros and M. Crovella, “Characteristics of WWW client-based trace”, Technical report TR-95-010, Boston University, Department of Computer Science, 1995.

[59] E. Dahlman, B. Gudmundson, M. Nilsson and J. Sköld, “UMTS/IMT-2000 based on wideband CDMA”, *IEEE Communications magazine*, vol. 36, no. 9, pp. 70-80, 1998.

[60] D. Dardari, V. Tralli and R. Verdone, “On the capacity of slotted Aloha with Rayleigh fading: the role played by the number of interferers”, *IEEE Communication letters*, vol. 2, no. 5, pp. 155-157, 2000.

[61] 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*, Würzburg, Germany, 2002.

[62] N.M. van Dijk, *“Queueing networks and product forms: a systems approach”*, John Wiley & Sons, New York, USA, 1993.

[63] N. Dimitriou, R. Tafazolli and G. Sfikas, “Quality of service for multimedia CDMA”, *IEEE Communications magazine*, vol. 38, no. 7, pp 88-94, 2000.

[64] C. Douligeris, “Multiobjective flow control in telecommunication networks”, *Proceedings of INFOCOM ’92*, Florence, Italy, 1992.

[65] J.J. Egli, “Radio propagation above 40 MC/s over irregular terrain”, *Proceedings of the IRE*, vol. 45, pp. 1383-1391, 1957.

[66] U. Ehrenberger and K. Leibnitz, “Impact of clustered traffic distributions in CDMA radio network planning, *Proceedings of ITC 16*, Edinburgh, Scotland, pp. 129-138, 1999.

[67] A.K. Erlang, “Sandsynlighedsregning og Telefonsamtaler” (in Danish; translated: “The theory of probabilities and telephone conversations”), *Nyt tidsskrift for matematik B*, vol. 20, pp. 33-39, 1909.

[68] V.M. Espinosa Velez, L. Jorguseski, R. Litjens, E.R. Fledderus, R. Prasad, “Downlink radio resource estimation and control in WCDMA cellular system with voice and data users”, *Proceedings of WPMC ’01,* Aalborg, Denmark, 2001.

[69] ETSI GSM 02.34, *“Digital cellular telecommunications systems (phase 2+),* H*igh* S*peed* C*ircuit* S*witched* D*ata, service description, stage 1”*, ETSI, France, 1997.

[70] ETSI GSM 02.60, *“Digital cellular telecommunications systems (phase 2+),* G*eneral* P*acket* R*adio* S*ervice, service description, stage 1”*, ETSI, France, 1999.

[71] ETSI GSM 03.34, *“Digital cellular telecommunications system (phase 2+);* H*igh* S*peed* C*ircuit* S*witched* D*ata, stage 2”,* ETSI, France, 1998.

[72] ETSI GSM 03.60, *“Digital cellular telecommunications systems (phase 2+),* G*eneral* P*acket* R*adio* S*ervice, service description, stage 2”*, ETSI, France, 1997.

[73] ETSI GSM 03.64, *“Overall description of the* GPRS *radio interface, stage 2”*, ETSI, France, 1998.

[74] ETSI GSM 04.60, *“Digital cellular telecommunications systems (phase 2+);* G*eneral* P*acket* R*adio* S*ervice* (GPRS*);* M*obile* S*tation* (MS*)* - B*ase* S*tation* S*ystem* (BSS*) interface;* R*adio* L*ink* C*ontrol /* M*edium* A*ccess* C*ontrol* (RLC/MAC*) protocol”*, ETSI, France, 2000.

[75] ETSI GSM 04.64, *“Logical link control* (LLC*) layer specification”*, v5.1.0, ETSI, France, 1997.

[76] ETSI GSM 04.65, *“*M*obile* S*tation* (MS*)* - S*erving* GPRS S*upport* N*ode* (SGSN*);* S*ubnetwork* D*ependent* C*onvergence* P*rotocol* (SNDCP*)”*, ETSI, France, 1997.

[77] ETSI GSM 08.08, *“Digital cellular telecommunications system (phase 2); Mobile-services Switching Centre - Base Station System (MSC-BSS) interface; layer 3 specification”,* ETSI, France, 1998.

[78] ETSI UMTS 30.04, *"UMTS terrestrial radio access; concept evaluation"*, ETSI, France, 1997.

[79] J.S. Evans and D. Everitt, "Effective bandwidth-based admission control for multi-service CDMA cellular networks", *IEEE Transactions on vehicular technology*, vol. 48, no. 1, pp. 36-46, 1999.

[80] S.N. Fabri, S. Worral, A. Sadka and A. Kondoz, "Real-time video communications over GPRS", *Proceedings of 3G Mobile communication technologies*, London, UK, pp. 426-430, 2000.

[81] G. Fayolle, I. Mitrani and R. Iasnogorodski, "Sharing a Processor Among Many Jobs", *Journal of the Association for Computing Machinery*, vol. 27, no. 3, pp. 519-532, 1980.

[82] C.H. Foh, B. Meini, B. Wydrowski and M. Zukerman, "Modelling and performance evaluation of GPRS", *Proceedings of VTC '01*, Rhodos, Greece, 2001.

[83] C.H. Foh and M. Zukerman, "Performance analysis of the IEEE 802.11 MAC protocol", *Proceedings of European Wireless '02*, Florence, Italy, 2002.

[84] G.J. Foschini, B. Gopinath and Z. Miljanic, "Channel cost of mobility", *IEEE Transactions on vehicular technology,* vol. 42, no. 4, pp. 414-424, 1993.

[85] G.J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence", *IEEE Transactions on vehicular technology,* vol. 42, no. 4, pp. 641-646, 1993.

[86] A. Furuskär, M. Frodigh, H. Olofsson and J. Sköld, "System performance of EDGE, a proposal for enhanced data rates in existing digital cellular systems", *Proceedings of VTC '98*, Ottawa, Canada, pp. 1284-1289, 1998.

[87] A. Furuskär, S. Mazur, F. Müller and H. Olofsson, "EDGE: enhanced data rates for GSM and TDMA/136 evolution", *IEEE Personal communications magazine*, pp. 56-66, June 1999.

[88] A. Furuskär, S. Parkvall, M. Persson and M. Samuelsson, "Performance of WCDMA high speed packet data", *Proceedings of VTC '02,* Birmingham, USA, 2002.

[89] J. Gardiner and B. West (editors), *"Personal communication systems and technologies"*, Artech House, Boston, USA, 1995.

[90] P.-P. Giesberts, Agere Systems, The Netherlands, *private communications*, 2003.

[91] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver and C.E. Wheatley, "On the capacity of a cellular CDMA system", *IEEE Transactions on vehicular technology*, vol. 40, no. 2, pp. 303-312, 1991.

[92] H. Granbohm and J. Wiklund, "GPRS - General Packet Radio Service", *Ericsson review*, no. 2, pp. 82-88, 1999.

[93] GSM association ISG, *"Typical radio parameter sets"*, v1.3, 2000.

[94] GSM association, "GSM *statistics*", available at `http://www.gsmworld.com/news/statistics/index.shtml`.

[95] Z. Hadzi-Velkov, Sts. Cyril and Methodius University, Skopje, Macedonia, *private communications*, 2003.

[96] Z. Hadzi-Velkov and B. Spasenovski, "Capture effect in IEEE 802.11 wireless LANs", *Proceedings of IEEE ICWLHN '01*, Singapore, 2001.

[97] Z. Hadzi-Velkov and B. Spasenovski, "IEEE 802.11 DCF with capture over Ricean-fading channel", *presentation given at the third IEEE Workshop on WLANs '01*, Boston, USA, 2001, available at `http://www.wlan01.wpi.edu/proceedings/wlan26d.pdf`.

[98] D. Harchol-Balter and A. Downey, "Exploiting process lifetime distributions for dynamic load balancing", *Proceedings of SIGMETRICS '96*, Philadelphia, USA, pp. 13-24, 1996.

[99] N. Hegde, *"On the impact of mobility on the performance of modern wireless networks",* Ph.D. thesis, University of Missouri, Kansas City, USA, 2000.

[100] K.W. Helmersson and G. Bark, "Performance of downlink shared channels in WCDMA radio networks", *Proceedings of VTC '01,* Rhodes, Greece, 2001.

[101] R.W. Hendriks, KPN Mobile, The Netherlands, *private communications*, 1998.

[102] A. Heras-Brandin, P. Bartolomé-Pascual, D. Gómez-Mateo and J. Izquierdo-Arce, "A multiservice dimensioning procedure for 3G CDMA", *Proceedings of 3G Mobile communication technologies '00*, London, UK, pp. 406-409, 2000.

[103] T.S. Ho and K.C. Chen, "Performance Analysis of IEEE 802.11 CSMA/CA Medium Access Control protocol", *Proceedings of IEEE PIMRC '96*, Taipei, Taiwan, pp. 407-411, 1996.

[104] S. Hoff, M. Meyer and A. Schieder, "A performance evaluation of Internet access via the General Packet Radio Service of GSM", *Proceedings of VTC '98*, Ottawa, Canada, pp. 1760-1764, 1998.

[105] H. Holma and A. Toskala (editors), "WCDMA *for* UMTS*: radio access for third generation mobile communications*", John Wiley & Sons, Chichester, England, 2002.

[106] G.J. Holtzmann and B. Pehrson, *"The early history of data networks"*, IEEE Computer Society Press, Los Alamitos, USA, 1995.

[107] J.M. Holtzman, "CDMA forward link waterfilling power control", *Proceedings of VTC '00,* Tokyo, Japan, 2000.

[108] D. Hong and S.S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures", *IEEE Transactions on vehicular technology*, vol. 35, no. 3, pp. 77-92, 1986. See also CEAS technical report, no. 773, College of Engineering and Applied Sciences, State University of New York, Stony Brook, NY 11794, USA.

[109] H. Honkasalo, K. Pehkonen, M.T. Niemi and A.T. Leino, "WCDMA and WLAN for 3G and beyond", *IEEE Wireless communications magazine*, vol. 9, no. 12, pp. 14-18, 2002.

[110] IEEE, "Wireless LAN Medium Access Control (MAC) and PHYsical layer (PHY) specifications", P802.11, 1997.

[111] IEEE, "Supplement to standard for telecommunications and information exchange between systems - LAN/MAN specific requirements - part 11: wireless Medium Access Control (MAC) and PHYsical layer (PHY) specifications: higher speed physical layer extension in the 2.4 GHz band", P802.11B/D7.0, 1999.

[112] IEEE, "Draft supplement to standard for telecommunications and information exchange between systems - LAN/MAN specific requirements - part 11: wireless Medium Access Control (MAC) and PHYsical layer (PHY) specifications: enhancements for Quality Of Service (QOS)", P802.11E/D1, 2001.

[113] IEEE, "Draft recommended practice for multi-vendor access point interoperability via an inter-access point protocol across distribution systems supporting IEEE 802.11 operation", Std 802.11F/D1, 2001.

[114] S. Irons, C. Johnson, A. King and D. McFarlane, "Supporting the successful deployment of third generation public cellular technologies – system dimensioning and network planning", *Proceedings of 3G mobile communication technologies '00*, London, UK, pp. 156-160, 2000.

[115] Y. Ishikawa and N. Umeda, "Capacity design and performance of call admission control in cellular CDMA systems", *IEEE Journal on selected areas in communications*, vol. 15, no. 8, pp. 1627-1635, 1997.

[116] R. Jäntti and S.-L. Kim, "Transmission rate scheduling for the non-real time data in a cellular CDMA system", *IEEE Communications Letters,* vol. 5, no. 5, pp. 200 -202, 2001.

[117] P.R. Jelenković, A.A. Lazar and N. Semret, "The effect of multiple time scales and subexponentiality in MPEG video streams on queueing behavior", *IEEE Journal on selected areas in communications*, vol. 15, no. 6, pp. 1052-1071,

1997.

[118] P.R. Jelenković and P. Momčilović, "Resource sharing with subexponential distributions", *Proceedings of INFOCOM '02*, New York, USA, 2002.

[119] J.-Y. Jeng, C.-W. Lin and Y.-B. Lin, "Dynamic scheduling for GSM data services", *IEICE Transactions on communications*, vol. E80, no. B(2), pp. 296-300, 1997.

[120] C. Johansson, L. de Verdier and F. Khan, "Performance of different scheduling strategies in a packet radio system", *Proceedings of ICUPC '98*, Florence, Italy, pp. 267-271, 1998.

[121] N. Joshi, S.R. Kadaba, S. Patel and G.S. Sundaram, "Downlink scheduling in CDMA data networks", *Proceedings of MOBICOM '00,* Boston, USA, pp. 179-190, 2000.

[122] R. Kalden, I. Meirick and M. Meyer, "Wireless Internet access based on GPRS", *IEEE Personal communications magazine*, pp. 8-18, 2000.

[123] J. Kalliokulju, "Quality of service management functions in 3rd generation mobile telecommunication networks", *Proceedings of WCNC '99*, New Orleans, USA, pp. 1283-1287, 1999.

[124] J.S. Kaufman, "Blocking in a shared resource environment", *IEEE Transactions on communications*, vol. 29, no. 10, pp. 1474-1481, 1981.

[125] F.P. Kelly, *"Reversibility and stochastic networks"*, John Wiley & Sons, New York, USA, 1979.

[126] K.D. Kennedy and R. Litjens, "Performance evaluation of a hybrid radio resource allocation algorithm in a GSM/GPRS network", *Proceedings of PIMRC '99*, Osaka, Japan, pp. 131-136, 1999.

[127] A.A. Kherani and A. Kumar, "Performance analysis of TCP with nonpersistent sessions", *Proceedings of the Workshop on Modelling of flow and congestion control*, Paris, France, 2000.

[128] A.A. Kherani and A. Kumar, "Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet", *Proceedings of INFOCOM '02*, New York, USA, 2002.

[129] S.-L. Kim, Z. Rosberg and J. Zander, "Combined power control and transmission rate selection in cellular networks", *Proceedings of VTC '99,* Amsterdam, The Netherlands, pp. 1653-1657, 1999.

[130] L. Kleinrock, "Analysis of a time-shared processor", *Naval research logisitics quarterly*, vol. 11, pp. 59-73, 1964.

[131] L. Kleinrock, "Time-shared systems: a theoretical treatment", *Journal of the Association for Computing Machinery*, vol. 14, no. 2, pp. 242-261, 1967.

[132] L. Kleinrock, *"Queueing systems, volume II"*, John Wiley & Sons, New York, USA, 1976.

[133] J. Knutsson, P. Butovitsch, M. Persson, and R.D. Yates, "Downlink admission control strategies for CDMA systems in a Manhattan environment", *Proceedings of VTC '98*, Ottawa, Canada, pp. 1453-1457, 1998.

[134] C. Konstantinopoulou, K. Koutsopoulos, P. Demestichas and M. Theologou, "Performance of a multi-service GSM/GPRS-capable network under various loading conditions", *Proceedings of VTC '01*, Rhodos, Greece, 2001.

[135] D.M. Kreps, *"A course in microeconomic theory"*, Harvester Wheatsheaf, Hertfordshire, United Kingdom, 1990.

[136] D. Kristic and L.M. Correia, "Influence of mobility from handover in cellular planning", *Proceedings of PIMRC '99*, Osaka, Japan, pp. 126-130, 1999.

[137] K. Kumaran and P. Whiting, "Rate processor sharing: a robust technique for scheduling data transmissions in CDMA wireless networks", *Proceedings of Multiaccess, mobility and teletraffic for wireless communications,* Venice, Italy, 1999.

[138] J. Laiho, A. Wacker and T. Novosad (editors), *"Radio network planning and optimisation for* UMTS*"*, John Wiley & Sons, Chichester, England, 2001.

[139] P. Lassila, J.L. van den Berg, M.Mandjes and R.E. Kooij, "An integrated packet/flow level model for TCP performance analysis", *Proceedings of ITC 18*, Berlin, Germany, 2003.

[140] C.T. Lau and C. Leung, "Capture models for mobile packet radio networks", *IEEE Transactions on Communications*, vol. 40, no. 5, pp. 917-925, 1992.

[141] W.C.Y. Lee,. *"Mobile communications design fundamentals"*, Howard W. Sams & Co., Indianapolis, USA, 1986.

[142] W.C.Y. Lee, *"Mobile cellular telecommunications – analog and digital systems"*, McGraw-Hill, New York, USA, 1995.

[143] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "On the self-similar nature of Ethernet traffic", *Proceedings of SIGCOMM '93*, San Francisco, USA, pp. 183-193, 1993.

[144] W. Li and A.S. Alfa, "Channel reservation for handoff calls in a PCS network", *IEEE Transactions on vehicular technology*, vol. 49, no. 1, pp. 95-104, 2000.

[145] J.-P. Linnartz, *"Narrowband land-mobile radio networks"*, Artech House, Boston, USA, 1993.

[146] D. Lister, S. Dehghan, R. Owen and P. Jones, "UMTS capacity and planning issues", *Proceedings of 3G Mobile communication technologies '00*, London, UK, pp. 218-223, 2000.

[147] M.J.J. Litjens, *"Engineering of lipase-catalysed conversions in organic solvents containing ammoniun salts"*, Ph.D. thesis, Technische Universiteit Delft, The Netherlands, 2000.

[148] R. Litjens, "The impact of mobility on UMTS network planning", *Proceeding of VTC '01,* Rhodes, Greece, 2001.

[149] R. Litjens, "The impact of mobility on UMTS network planning", *Computer networks,* vol. 38, no. 4, 2002.

[150] R. Litjens and J.L. van den Berg, "Performance analysis of adaptive scheduling in integrated services UMTS networks", *Proceedings of MWCN '02,* Stockholm, Sweden, pp. 3-7, 2002.

[151] R. Litjens and R.J. Boucherie, "Radio resource sharing in a GSM/GPRS network", *Proceedings of the ITC specialist seminar on Mobile systems and mobility*, Lillehammer, Norway, pp. 261-274, 2000.

[152] R. Litjens and R.J. Boucherie, "Evolutie van mobiele cellulaire telecommunicatie-netwerken - parallele evolutie van netwerk-technologie en wiskundige performance evaluatie-modellen", *Proceedings of Symposium Wiskunde Toegepast*, Maastricht, The Netherlands, 2000.

[153] R. Litjens and R.J. Boucherie, "Performance analysis of fair channel sharing policies in an integrated cellular voice/data network", *Telecommunication systems*, vol. 19, no. 2, pp. 147-186, 2002.

[154] R. Litjens and R.J. Boucherie, "Elastic calls in an integrated services network: the greater the call size variability the better the Quality-of-Service", *Performance evaluation,* vol. 52, no. 4, pp. 193-220, 2003.

[155] R.Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie and M. Fleuren, "Performance analysis of WLANs: an integrated packet/flow level approach", *Proceedings of ITC 18*, Berlin, Germany, 2003.

[156] I. López, P.J. Ameigeiras, J. Wigard and P. Mogensen, "Downlink radio resource management for IP packet services in UMTS", *Proceedings of VTC '01,* Rhodes, Greece, 2001.

[157] R. Love, A. Ghosh, R. Nikides, L. Jalloul, M. Cudak and B. Classon, "High speed downlink packet access performance", *Proceedings of VTC '01,* Rhodes, Greece, 2001.

[158] Y. Lu and R.W. Brodersen, “Integrating power control, error correction coding, and scheduling for a CDMA downlink system”, *IEEE Journal on selected areas in communications,* vol. 17, no. 5, pp. 978-989, 1999.

[159] R.C.V. Macario (editor), *“Personal & mobile radio systems”*, Peter Peregrinus, London, United Kingdom, 1991.

[160] V.H. MacDonald, “The cellular concept”, *Bell system technical journal*, vol. 58, no. 1, pp. 15-43, 1979.

[161] M.R.H. Mandjes and M.J.G. van Uitert, “Transient analysis of traffic generated by bursty sources, and its application to measurement-based admission control”, *Telecommunication systems*, vol. 15, no. 3, pp. 295-321, 2000.

[162] S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz and L. Stibor, “IEEE 802.11E wireless LAN for quality of service”, *Proceedings of European Wireless ’02*, Florence, Italy, 2002.

[163] Marconi corporation, “**marconi**calling”, available at `http://www.marconi.com/html/about/marconihistory.htm`.

[164] L. Massoulié and J.W. Roberts, “Arguments in favour of admission control for TCP flows”, *Proceedings of ITC 16*, Edinburgh, Scotland, 1999.

[165] D. McMillan, “Traffic modelling and analysis for cellular mobile networks”, *Proceedings of ITC 13*, Copenhagen, Denmark, pp. 627-632, 1991.

[166] A. Mehrotra, *“Cellular radio – analog and digital systems”*, Artech House, Boston, USA, 1994.

[167] N.B. Mehta, L. Greenstein, T. Willis and Z. Kostic, “Analysis and results for the orthogonality factor in WCDMA downlinks”, *Proceedings of VTC ’02,* Birmingham, USA, 2002.

[168] R.D. van der Mei, J.L. van den Berg, R. Vranken and B.M.M. Gijsen, “Sojourn times in multiple-server processor sharing systems with priorities”, to appear in *Performance evaluation,* 2003.

[169] M. Meyer, “TCP performance over GPRS”, *Proceedings of WCNC ’99,* New Orleans, USA, pp. 1248-1252, 1999.

[170] D. Mitra, “An asynchronous distributed algorithm for power control in cellular radio systems”, *Proceedings of the 4th WinLab workshop on 3rd Generation wireless information networks,* Piscataway, USA, pp. 249–257, 1993.

[171] M. Mouly and M.-B. Pautet, *“The GSM system for mobile communications”*, published by the authors, France, 1992.

[172] M.F. Neuts, *“Matrix-geometric solutions in stochastic models: an algorithmic approach”*, Johns Hopkins University Press, Baltimore, USA, 1981.

[173] R. Núñez Queija, *"Processor-sharing models for integrated-services networks",* Ph.D. thesis, Technische Universiteit Eindhoven, The Netherlands, 2000.

[174] R. Núñez Queija, "Sojourn times in non-homogeneous QBD processes with processor sharing", *Stochastic models,* vol. 17, pp. 61-92, 2001.

[175] R. Núñez Queija, J.L. van den Berg, and M.R.H. Mandjes, "Performance evaluation of strategies for integration of elastic and stream traffic", *Proceedings of ITC 16*, Edinburgh, Scotland, pp. 1039-1050, 1999.

[176] T. Ojanperä and R. Prasad, "An overview of air interface multiple access for IMT-2000/UMTS", *IEEE Communications magazine*, vol. 36, no. 9, pp. 82-95, 1998.

[177] M. Ostrowski, "Efficient transmission of integrated voice and data in wireless networks", *Proceedings of ICC/SUPERCOMM '96*, Dallas, USA, pp. 721-727, 1996.

[178] T. Ott, "The sojourn time distribution in the $M/G/1$ queue with processor sharing", *Journal of applied probability*, vol. 21, pp. 360-378, 1984.

[179] D.L. Pallant and P.G. Taylor, "Modeling handovers in cellular mobile networks with dynamic channel allocation", *Operations research*, vol. 43, no. 1, pp. 33-42, 1995.

[180] Q. Pang, A. Bigloo, V.C.M. Leung and C. Scholefield, "Service scheduling for General Packet Radio Service classes", *Proceedings of WCNC '99*, New Orleans, USA, pp. 1229-1233, 1999.

[181] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case", *Proceedings of INFOCOM '92*, Florence, Italy, pp. 915-924, 1992.

[182] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple node case", *Proceedings of INFOCOM '93*, San Francisco, USA, pp. 521-530, 1993.

[183] K. Park and W. Willinger, "Self-similar network traffic: an overview", Chapter 1 of *"Self-similar network traffic and performance evaluation"* (eds.: K. Park and W. Willinger), John Wiley & Sons, Chichester, England, 2000.

[184] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, M. Persson, "The evolution of WCDMA towards higher speed downlink packet data access", *Proceedings of VTC '01,* Rhodes, Greece, 2001.

[185] S. Parkvall, J. Peisa, A. Furuskär, M. Samuelsson and M. Persson, "Evolving WCDMA for improved high speed mobile Internet", *Proceedings of the Future Telecommunications Conference '01,* Beijing, China, 2001.

[186] C. Parsa and J.J. Garcia-Luna-Aceves, "Improving TCP performance over wireless networks at the link layer", *Mobile networks and applications*, vol. 5, pp. 57-71, 2000.

[187] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modelling", *IEEE/ACM Transactions on networking*, vol. 3, no. 3, pp. 226-244, 1995.

[188] R. Prasad, W. Mohr and W. Konhäuser (editors), *"Third generation mobile communications systems"*, Artech House, Boston, USA, 2000.

[189] N. Prasad and A. Prasad (editors), *"WLAN systems and wireless IP for next generation communications"*, Artech House, Norwood, USA, 2002.

[190] S. Rácz, B.P. Gerö and G. Fodor, "Flow level performance analysis of a multiservice system supporting elastic and adaptive services", *Performance evaluation,* vol. 49, no. 1/4, pp. 451-469, 2002.

[191] S. Ramakrishna and J.M. Holtzman, "A scheme for throughput maximization in a dual-class CDMA system", *Proceedings of ICUPC '97*, pp. 623-627, 1997.

[192] S.M. Redl, M.K. Weber and M.W. Oliphant, *"An introduction to GSM"*, Artech House, Boston, USA, 1995.

[193] J.M. Rivadeneyra Sicilia and J. Miguel-Alonso, "A communication architecture to access data services through GSM", *Proceedings of INDC '98*, Aveiro, Portugal, pp. 1-11, 1998.

[194] J.W. Roberts, "A service system with heterogeneous user requirements - application to multiservice telecommunication systems", *Performance of data communications and their applications*, G. Pujolle (editor), North Holland, New York, USA, 1981.

[195] J.W. Roberts et al. (eds.), *"Broadband network teletraffic (COST 242)"*, Springer-Verlag, Berlin-Heidelberg, Germany, 1996.

[196] J.W. Roberts, "Quality of service guarantees and charging in multiservice networks", *IEICE Transactions on communications*, vol. E81, no. B(5), 1998.

[197] J.W. Roberts and L. Massoulié,. "Bandwidth sharing and admission control for elastic traffic", *Proceedings of the ITC specialist seminar on Teletraffic issues related to multimedia and nomadic communications*, Yokohama, Japan, 1998.

[198] K.W. Ross, *"Multiservice Loss Models for Broadband Telecommunication Networks"*, Springer-Verlag, London, United Kingdom, 1995.

[199] S.M. Ross, *"Stochastic processes"*, John Wiley & Sons, New York, USA, 1983.

[200] S.M. Ross, *"A first course in probability"*, Macmillan Publishing Company, New York, USA, 1984.

[201] K. Ruttik, *"Radio resource management"*, lecture notes, available at `http://www.comlab.hut.fi/opetus/238/lecture10_RRM.pdf`, Helsinki University of Technology, Finland, 2002.

[202] J. Sachs, T. Balon and M. Meyer, "Congestion control in WCDMA with respect to different service classes", *Proceedings of European Wireless '99,* Munich, Germany, pp. 303-308, 1999.

[203] M. Sakata, S. Noguchi and J. Oizumi, "Analysis of a processor-based queueing model for time-sharing systems", *Proceedings of the 2nd Hawaii international conference on System sciences*, USA, pp. 625-628, 1969.

[204] M. Sakata, S. Noguchi and J. Oizumi, "An analysis of the $M/G/1$ queue under round robin scheduling", *Operations research*, vol. 19, pp. 371-385, 1971.

[205] J. Sau and C. Scholefield, "Scheduling and quality of service in the General Packet Radio Service", *Proceedings of ICUPC '98*, Florence, Italy, pp. 1067-1071, 1998.

[206] A. Schieder, U. Horn and R. Kalden, "Performance analysis of realtime applications in mobile packet switched networks", *Proceedings of European Wireless '99*, Munich, Germany, 1999.

[207] J.R. Schmidt, TNO Telecom, The Netherlands, *private communications*, 2003.

[208] E. Seneta, *"Non-negative matrices"*, Springer-Verlag, New York, USA, 1981.

[209] R. Serfozo, *"Introduction to stochastic networks"*, Springer-Verlag, New York, USA, 1999.

[210] W.R. Stevens, *"TCP/IP illustrated, Volume 1: The protocols"*, Addison-Wesley, Reading, USA, 1994

[211] D.J. Struik (editor), *"A source book in mathematics 1200-1800"*, Princeton University Press, Princeton, USA, pp. 89-93, 1986.

[212] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks", *IEEE Communications magazine*, vol. 29, no. 11, pp. 42-46, 1991.

[213] H.C. Tijms, *"Stochastic modelling and analysis: a computational approach"*, John Wiley & Sons, Chichester, England, 1986.

[214] B. Tsybakov and N.D. Georganas, "Self-similar processes in communications networks", *IEEE Transactions on information theory*, vol. 44, no. 5, pp. 1713-1725, 1998.

[215] D. Turina, P. Beming, E. Schoster and A. Andersson, "A proposal for multi-slot MAC layer operation for packet data channel in GSM", *Proceedings of ICUPC '96*, Cambridge, USA, pp. 572-576, 1996.

[216] UMTS forum, MobilenniuM newsletter, available at `http://www.umts-forum.org/MobilleniuM/feb03/mobilennium/frames.htm`, February, 2003.

[217] F.J. Velez and L.M. Correia, "Traffic from mobility in mobile broadband systems", *Telektronikk*, vol. 3/4, pp. 95-101, 1998.

[218] A.J. Viterbi, *"CDMA: principles of spread spectrum communication",* Addison-Wesley, Reading, USA, 1995.

[219] E.Th. de Vries and M. Aldén, "Traffic modelling", Chapter 2 of *"Generic radio transport"* (eds.: S. Grujev and R. Hekmat), Research report R&D-RA-97-958, KPN Research and Telia Research, 1997.

[220] A.R. Ward and W. Whitt, "Predicting response times in processor-sharing queues", *Proceedings of the Fields Institute conference on Communication networks*, Providence, USA, pp. 1-29, 2000.

[221] C. Ware, J.F.Chicharo and T.Wysocki, "Modelling capture behaviour in IEEE 802.11 radio modems", *Proceedings of IEEE VTC '01*, Atlantic City, NJ, 2001.

[222] J. Weinmiller, M. Schlager, A. Festag and A. Wolisz, "Performance study of access control in wireless LANs IEEE 802.11 DFWMAC and ETSI RES 10 HIPERLAN", *Mobile networks and applications*, vol. 2, no. 1, pp. 55-67, 1997.

[223] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level", *Proceedings of SIGCOMM '95*, Cambridge, USA, pp. 100-113, 1995.

[224] R.W. Wolff, *"Stochastic modeling and the theory of queues"*, Prentice-Hall, Englewood Cliffs, USA, 1989.

[225] H. Wu, Y. Peng, K. Long, S. Cheng and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement", *Proceedings of IEEE INFOCOM '02*, New York, USA, 2002.

[226] S.F. Yashkov, "A derivation of response time distribution for a $M/G/1$ processor sharing queue", *Problems in control and information theory*, vol. 12, pp. 133-148, 1983.

[227] S.F. Yashkov, "Processor sharing queues: some progress in analysis", *Queueing systems*, vol. 2, pp. 1-17, 1987.

[228] S.F. Yashkov, "Mathematical problems in the theory of processor sharing queueing systems", *Journal of Soviet mathematics*, vol. 58, pp. 101-147, 1992.

[229] C.H. Yoon and C.K. Un, "Performance of personal portable radio telephone systems with and without guard channels", *IEEE Journal on selected areas in communications*, vol. 11, no. 6, pp. 911-917, 1993.

[230] W. Yue and Y. Matsumoto, "An exact analysis for CSMA/CA protocol integrated voice/data wireless LANs", *Proceedings of IEEE GLOBECOM '00*, San Francisco, USA, 2000.

[231] J. Zander and Olav Queseth, *"Radio resource management for wireless networks"*, Artech House, Boston, USA, 2001.

[232] A.P. Zwart, *"Queueing systems with heavy tails"*, Ph.D. thesis, Technische Universiteit Eindhoven, The Netherlands, 2001.

[233] A.P. Zwart and O.J. Boxma, "Sojourn time asymptotics in the $M/G/1$ processor sharing queue", *Queueing systems,* vol. 35, pp. 141-166, 2000.

BIBLIOGRAPHY

[1] V.M. Espinosa Velez, L. Jorguseski, R. Litjens, E.R. Fledderus, R. Prasad, "Downlink radio resource estimation and control in WCDMA cellular system with voice and data users", *Proceedings of WPMC '01,* Aalborg, Denmark, 2001.

[2] K.D. Kennedy and R. Litjens, "Performance evaluation of a hybrid radio resource allocation algorithm in a GSM/GPRS network", *Proceedings of PIMRC '99*, Osaka, Japan, pp. 131-136, 1999.

[3] R. Litjens, "The impact of mobility on UMTS network planning", *Proceeding of VTC '01,* Rhodes, Greece, 2001.

[4] R. Litjens, "The impact of mobility on UMTS network planning", *Computer networks,* vol. 38, no. 4, 2002.

[5] R. Litjens and J.L. van den Berg, "Fair adaptive scheduling in integrated services UMTS networks", *submitted.*

[6] R. Litjens and J.L. van den Berg, "Performance analysis of adaptive scheduling in integrated services UMTS networks", *Proceedings of MWCN '02,* Stockholm, Sweden, pp. 3-7, 2002.

[7] R. Litjens and R.J. Boucherie, "Radio resource sharing in a GSM/GPRS network", *Proceedings of the ITC specialist seminar on Mobile systems and mobility*, Lillehammer, Norway, pp. 261-274, 2000.

[8] R. Litjens and R.J. Boucherie, "Evolutie van mobiele cellulaire telecommunicatienetwerken - parallele evolutie van netwerk-technologie en wiskundige performance evaluatie-modellen", *Proceedings of Symposium Wiskunde Toegepast*, Maastricht, The Netherlands, 2000.

[9] R. Litjens and R.J. Boucherie, "Quality-of-service differentiation in an integrated services GSM/GPRS network", *submitted,* 2001.

[10] R. Litjens and R.J. Boucherie, "Performance analysis of fair channel sharing policies in an integrated cellular voice/data network", *Telecommunication systems,* vol. 19, no. 2, pp. 147-186, 2002.

[11] R. Litjens and R.J. Boucherie, “Elastic calls in an integrated services network: the greater the call size variability the better the Quality-Of-Service”, *Performance evaluation*, vol. 52, no. 4, pp. 193-220, 2003.

[12] R. Litjens and R.J. Boucherie, “Performance analysis of Downlink Shared CHannels in a UMTS network”, *submitted*, 2002.

[13] R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie and M. Fleuren, “Performance analysis of WLANs: an integrated packet/flow level approach”, *Proceedings of ITC 18*, Berlin, Germany, 2003.

[14] R. Litjens, J.L. van den Berg and R.J. Boucherie, “Throughput measures for processor sharing models”, *submitted*, 2003.

ACRONYMS

3G	3rd Generation
3GPP	3rd Generation Partnership Project
16-QAM	Quadrature Amplitude Modulation
ABR	Available Bit Rate
ACK	ACKnowledgment
A-DCH	Associated Dedicated CHannel
AM	Amplitude Modulation
a.o.	among others
AP	Adaptive scheduling - Power fair
AP	Access Point
AR	Adaptive scheduling - Rate fair
ARQ	Automatic Repeat reQuest
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
BLER	BLock Error Rate
BSC	Base Station Controller
BSS	Basic Service Set
BTS	Base Transceiver Station
C/I	carrier-to-interference ratio
CAC	Call Admission Control
CBR	Constant Bit Rate
CDF	Cumulative Distribution Function

CDMA	Code Division Multiple Access
CEPT	Conférence Européenne des administrations des Postes et des Télécommunications
cf.	confer
CM	Capture Model
COST	COoperation in the field of Scientific and Technical research
CPCH	Common Packet CHannel
CS	Coding Scheme
CSD	Circuit-Switched Data
CSMA(/CA)	Carrier Sense Multiple Access (with Collision Avoidance)
CT	Corporate Technology
CTS	Clear To Send
dB	decibel
dBm	decibel relative to one milliWatt
DCA	Dynamic Channel Allocation
DCF	Distributed Coordination Function
DCH	Dedicated CHannel
DIFS	Distributed InterFrame Space
DPS	Discriminatory Processor Sharing
DSCH	Downlink Shared CHannel
E_b/N_o	energy-per-bit to interference-plus-noise density ratio
EDGE	Enhanced Data rates for Global Evolution
e.g.	exempli gratia
ETSI	European Telecommunications Standards Institute
FACH	Forward Access CHannel
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
FER	Frame Error Rate

FIFO	First-In First-Out
FM	Frequency Modulation
FP	Fixed scheduling - Power fairness
FR	Fixed scheduling - Rate fairness
GGSN	Gateway GPRS Support Node
GHz	GigaHertz
GI	General Independent
GMSC	Gateway MSC
GMSK	Gaussian Minimum Shift Keying
GOS	Grade Of Service
GPRS	General Packet Radio Service
GPS	Generalised Processor Sharing
GSM	Groupe Spécial Mobile
GSM	Global System for Mobile communications
h.l.	hoc loco
HSCSD	High-Speed Circuit-Switched Data
HSDPA	High-Speed Downlink Packet Access
HS-DSCH	High-Speed Downlink Shared CHannel
Hz	Hertz
i.e.	id est
IEEE	Institute of Electrical and Electronics Engineers
IMT-2000	International Mobile Telecommunications-2000
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISM	Industrial, Scientific, Medical
ISO	International Organisation for Standardisation
ITU	International Telecommunication Union
kbits/s	kilobits per second

kchips/s	kilochips per second
kHz	kiloHertz
km	kilometer
km/h	kilometer per hour
KPN	Koninklijke PTT Nederland
LAN	Local Area Network
LOS	Line Of Sight
m	meter
MAC	Medium Access Control
MAN	Metropolitan Area Network
Mbits/s	Megabits per second
MHz	MegaHertz
MIMO	Multiple Input Multiple Output
MPEG	Motion Pictures Expert Group
ms	millisecond(s)
MS	Mobile Station
MSC	Mobile Switching Center
NAV	Network Allocation Vector
NMT	Nordic Mobile Telephony
NTT	Nippon Telegraph & Telephone corporation
OVSF	Orthogonal Variable Spreading Factor
PAN	Personal Area Network
PASTA	Poisson Arrivals See Time Averages
PCF	Point Coordination Function
PCU	Packet Control Unit
PDN	Packet Data Network
PDP	Packet Data Protocol
PDF	Probability Density Function

PDP	Packet Data Protocol
PHY	PHYsical layer
PLMN	Public Land Mobile Network
PS	Processor Sharing
PSTN	Public Switched Telephone Network
QOS	Quality Of Service
QPSK	Quadrature Phase Shift Keying
R&D	Research and Development
RACH	Random Access CHannel
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RR	Round Robin
RRR	Radio Resource Reservation
RTS	Request To Send
SGSN	Serving GPRS Support Node
SHL	Speech/High-priority data/Low-priority data
SIFS	Short InterFrame Space
SMS	Short Message Service
STA	STAtion
STW	Stichting Technische Wetenschappen
S(V)(D)	Speech(/Video)(/Data)
SVHL	Speech/Video/High-priority data/Low-priority data
TCP	Transmission Control Protocol
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
TE	Terminal Equipment
TG	Task Group
TNO	Toegepast Natuurwetenschappelijk Onderzoek

TPC	Transmission Power Control
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
VBR	Variable Bit Rate
viz.	videlicet
WAN	Wide Area Network
WECA	Wireless Ethernet Compatibility Alliance
WIFI	WIreless FIdelity
WLAN	Wireless Local Area Network
WMC	Wireless and Mobile Communications
WRR	Weighted Round Robin
WWW	World Wide Web

SAMENVATTING

De primaire doelstelling van een netwerk-operator is de kosten-efficiënte planning en exploitatie van een netwerk, dat dienstaanbieders toestaat om de eindgebruiker adequate doch betaalbare kwaliteit aan te bieden. Dit streven wordt gekenmerkt door een inherente wisselwerking tussen netwerkcapaciteit en dienstkwaliteit. Een scala aan capaciteitsallocatie-mechanismen staat de netwerkoperator op verschillende tijdschalen ter beschikking om de gewenste balans tussen capaciteit en kwaliteit na te streven, die overeenkomt met zijn positioneringsstrategie in de draadloze telecommunicatiemarkt. Als voorbeeld, op een typische tijdschaal van weken richt *site planning* zich op het selecteren van geschikte locaties voor nieuw te plaatsen basisstations. Het *call admission control* mechanisme accepteert danwel weigert *call*-aanvragen op een tijdschaal van seconden, teneinde de dienstkwaliteit van zowel nieuwe als reeds opstaande *calls* te waarborgen. Het doel van *transmission power control* is om, op een tijdschaal van (milli)seconden, de zendvermogens te minimaliseren teneinde het interferentieniveau te beheersen en de kwaliteit van de verschillende radioverbindingen op een efficiënte wijze te waarborgen. Een inzichtelijk referentiemodel voor de uitgebreide suite van capaciteitsallocatie-mechanismen is opgenomen in Hoofdstuk 1.

De modellering en prestatie-analyse van capaciteitsallocatie in draadloze netwerken genereert essentiële inzichten in de wisselwerking tussen capaciteit en dienstkwaliteit, alsmede hoe deze wisselwerking optimaal bespeeld kan worden middels de verscheidene beschikbare mechanismen. Naarmate draadloze netwerken steeds complexer van aard worden door technologische ontwikkelingen en de aangeboden diensten steeds diverser worden in hun verkeerskarakteristieken en kwaliteitseisen, is een dergelijke modellering en analyse onbetwistbaar noodzakelijk. Dit proefschrift biedt waardevolle bijdragen op een drietal verschillende vlakken, met als doel om de invloed van de capaciteitsallocatie-mechanismen op de exploitatie en dienstverlening in cellulaire netwerken en draadloze *local area networks* (WLANs) te doorgronden en te optimaliseren. Gebaseerd op een solide basis van technologische kennis, worden allereerst de primaire aspecten van de beschouwde draadloze netwerken in werkbare modellen vervat. Vervolgens worden op maat gemaakte prestatie-analyse methoden ontwikkeld, die tenslotte worden uitgebuit om middels numerieke experimenten de inherente wisselwerking tussen capaciteit en kwaliteit bloot te leggen en te kwantificeren.

Hoofdstukken 2 tot en met 5 concentreren zich op de HSCSD- en GPRS-uitbreidingen van tweede-generatie GSM-netwerken. Vanuit het perspectief van prestatie-analyse van GSM/HSCSD/GPRS-netwerken staat het relatief afgelegen frequentie-hergebruik dat is opgelegd door de toegepaste *frequency-/time-division multiple access* technologie het toe om het onderzoek op zinvolle wijze te beperken tot een enkele cel.

Hoofdstuk 2 onderzoekt de integratie van spraaktelefonie en data-transmissie in een GSM/HSCSD-cel, waar data *calls* elastisch zijn in de zin dat ze zich flexibel kunnen aanpassen aan een variërende kanaaltoewijzing. Een algemeen kader van kanaaltoewijzingsregels wordt gepresenteerd, die bestaan uit een *call admission control* regel, een capaciteitsreservering voor een of beide diensten, een *rate control* regel die de aan data-transmissie toegewezen capaciteit dynamisch aanpast en een *processor sharing scheduling* discipline die de beschikbare capaciteit gelijkelijk verdeelt over de actieve data *calls*. Een extensieve Markov keten-analyse wordt toegepast om de prestatie van verschillende kanaaltoewijzingsregels te evalueren en te vergelijken. Naast een aantal prestatie-maten die direct uit de evenwichtsverdeling van de Markov keten kunnen worden afgeleid, zoals de dienstspecifieke blokkeringskansen, wordt tevens een analytische uitdrukking afgeleid voor de verwachte doorlooptijd van een toegelaten data *call*, geconditioneerd op zijn grootte en de systeemtoestand bij aankomst. Een dergelijke voorspelling van de doorlooptijd kan als een gewaardeerde dienst aan de databron worden teruggekoppeld.

Hoofdstuk 3 breidt het model en de analyse van Hoofdstuk 2 uit, teneinde de prestatie van een generiek GSM/GPRS-model met spraak, video, hoge en lage prioriteits-data-diensten te onderzoeken. Zowel de data als de nieuw geïntroduceerde video *calls* zijn elastisch van aard. Het kenmerkende verschil is, dat in het geval van de videodienst een meer genereuze kanaaltoewijzing en een dientengevolge hogere *throughput* de videokwaliteit verbetert zonder de houdtijd te beïnvloeden, terwijl voor data *calls* een hogere *throughput* een kortere doorlooptijd impliceert. Wederom wordt een uitvoerige Markov keten-analyse toegepast voor dit uitgebreide model, inclusief e.g. de conditionele verwachte video *throughput* als functie van de *call*-duur en de systeemtoestand bij aankomst.

Gezien de geobserveerde en algemeen erkende hoge variabiliteit van e.g. WWW pagina-groottes, laat Hoofdstuk 4 de exponentialiteitsaanname voor data *call* groottes los en presenteert een gevoeligheidsanalyse voor de verwachte doorlooptijden van data *calls* in een geïntegreerd GSM/GPRS-netwerk met spraak- en data-diensten. De opmerkelijke observatie wordt gepresenteerd en analytisch ondersteund dat in het beschouwde *processor sharing* model met variërende capaciteit (ten gevolge van de aankomst/vertrek-dynamiek van geprioritiseerd spraakverkeer), de verwachte doorlooptijden van data *calls* afnemen met een hogere variabiliteit van de data *call* groottes. Dit staat in contrast met e.g. *processor sharing* of *first-in first-out* wachtrijmodellen met vaste capaciteit, waarin ongevoeligheid respectievelijk zelfs het omgekeerde effect geldt. De invloed van de variabiliteit van de data *call* grootte op de dienstbeleving wordt verder onderzocht voor enkele uitgebreide modellen uit Hoofdstukken 2 en 3.

Als uitbreiding op de video *throughput* analyse van Hoofdstuk 3, presenteert Hoofdstuk 5 een meer theoretische analyse van *throughput* maten voor elastische calls in *processor sharing* systemen met vaste of random variërende capaciteit. Het eerstgenoemde alternatief modelleert e.g. een enkele GPRS-cel met louter elastisch (video of data) verkeer, terwijl het tweede alternatief correspondeert met een GSM/GPRS-cel die geprioritiseerde spraaktelefonie integreert met elastisch verkeer. Een aantal verschillende *throughput* maten voor elastische diensten wordt gedefinieerd, geanalyseerd en vergeleken, zoals de *call*-gemiddelde, de tijdsgemiddelde en de nieuw voorgestelde verwachte instantane *throughput*. Specifieke aandacht wordt besteed aan de gevoeligheid van de *throughput* maten voor de verdeling van de elastische *call*-groottes. De belangrijkste conclusie is dat de *call*-gemiddelde *throughput*, die het meest relevant wordt geacht vanuit het perspectief van de eindgebruiker, doch in het algemeen moeilijk te analyseren is, in alle beschouwde scenario's louter door de verwachte instantane *throughput* voortreffelijk benaderd wordt, die bovendien relatief eenvoudig kan worden afgeleid en behoorlijk ongevoelig is voor de verdeling van de elastische *call*-groottes.

Hoofdstukken 6 tot en met 8 concentreren zich op derde-generatie UMTS-netwerken. Het UMTS-radio netwerk is gebaseerd op *code-division multiple access* technologie, hetgeen een universeel frequentie-hergebruik impliceert, alsmede de inherente eigenschap dat de verbruikte capaciteit van een *call* afhangt van diens relatieve locatie. Dientengevolge behoort een zinvolle UMTS-studie zowel meerdere cellen als ook de willekeur van gebruikerslocatie te beschouwen, in tegensteling tot de GSM/HSCSD/GPRS-analyses van Hoofdstukken 2 tot en met 5. In feite zijn het daarmee precies die radio-aspecten die de systeemefficiëntie verhogen, die tevens de prestatie-analyse significant compliceren.

Hoofdstuk 6 richt zich op data-transmissie over D*ownlink* S*hared* CH*annels*, aantoonbaar het meest efficiënte UMTS transport kanaaltype om de verwachte *downlink* data-volumes te verwerken. In deze context beïnvloedt de dataverkeersbelasting de geleverde dienstkwaliteit op twee verschillende wijzen. Ten eerste impliceert een zwaardere dataverkeersbelasting een hevigere competitie voor DSCH capaciteit en derhalve langer transmissietijden. Ten tweede, aangezien iedere data *call* die op een DSCH wordt afgehandeld een zogenaamde A*ssociated* D*edicated* CH*annel* onderhoudt voor signaleringsdoeleinden, leidt een zwaardere dataverkeersbelasting tot een hoger interferentieniveau, een hogere foutkans voor verzonden pakketten en derhalve een lagere effectieve geaggregeerde DSCH *throughput*: des te hoger de transmissievraag, des te lager de geaggregeerde capaciteit. De algehele prestatie-analyse is ontleed in twee deelanalyses, teneinde de interferentie-aspecten van de verkeersdynamiek te scheiden. In de eerstgenoemde deelanalyse worden de effecten van gebruikerslocatie, *transmission power control* en interferentie op de effectieve DSCH *throughput* onderzocht. De laatstgenoemde deelanalyse vervat de dynamiek op *call*-niveau in een meerdimensionaal *processor sharing* model met toestandsafhankelijke effectieve transmissiesnelheden, die worden afgeleid van de resultaten uit de eerstgenoemde deelanalyse. Een reeks van geleidelijk meer gecompliceerde scenario's wordt geëvalueerd om de bovengenoemde invloeden van de dataverkeersbelasting op de ervaren dienstkwaliteit te demonstreren.

Hoofdstuk 7 richt zich op de efficiënte allocatie van capaciteit in een UMTS/HSDPA-netwerk met spraak- en data-diensten en streeft daarmee vergelijkbare doeleinden na als Hoofdstukken 2 en 3. Door de fundamenteel verschillende radio-technologieën verschillen de benodigde capaciteitsallocatie-handelingen echter aanmerkelijk. Een gemeenschappelijke eigenschap van de voor GSM/HSCSD/GPRS-netwerken bestudeerde dienstintegratie-strategieën is de eerlijke verdeling van de beschikbare capaciteit over de aanwezige data *calls*. Zoals in Hoofdstuk 7 wordt gedemonstreerd, kan deze *fairness* doelstelling niet simpelweg worden doorgetrokken naar een op *code-division multiple access* technologie gebaseerd radio netwerk, aangezien in dergelijke netwerken een essentieel verschil bestaat tussen *fairness* van gealloceerde capaciteit en *fairness* in termen van ervaren *throughput*. De efficiëntie en *fairness* doelstellingen worden nagestreefd middels *adaptive scheduling*, dat bestaat uit een gebalanceerde inzet van capaciteitsallocatie-mechanismen op verschillende tijdschalen. Op een tijdschaal van seconden tot minuten exploiteert *rate control* de inherente flexibiliteit en vertragingstolerantie van data *calls*, door de aan dataverkeer gealloceerde capaciteit te verlagen in tijden van een hoge spraakverkeersbelasting om de spraakkwaliteit te waarborgen, en deze te verhogen bij een lage spraakverkeersbelasting teneinde de data *throughputs* te bevorderen. Op een tijdschaal van (milli)seconden opereert de *packet scheduler* in overeenstemming met de geselecteerde *fairness* doelstelling. De prestaties van de *adaptive scheduling* algoritmen worden geëvalueerd middels analytische optimalisatie in combinatie met Monte Carlo simulaties. De numerieke experimenten demonstreren dat het mogelijk is om door middel van *adaptive scheduling* de kwaliteit van zowel de spraak- als data-diensten te verhogen, waarbij de meest significante prestatieverbetering door het geprioritiseerde spraakverkeer wordt ervaren.

Hoofdstuk 8 onderzoekt de invloed van mobiliteit op UMTS radio netwerk-planning, de geleverde dienstkwaliteit en de benodigde investeringskosten, die voortkomt uit twee afzonderlijke aspecten. Aan de ene kant impliceert een hogere mobiliteitsgraad strengere eisen voor de signaal-ruis-verhouding, ten gevolge van de gezamenlijke effecten van *multipath* propagatie, Doppler verschuivingen en *transmission power control* imperfecties. Dientengevolge dienen cel-groottes gereduceerd te worden om een zelfde capaciteit te handhaven. Aan de andere kant vereist een hogere mobiliteitsgraad een ruimere capaciteitsreservering ter ondersteuning van *call handovers*, teneinde de kans op vroegtijdige *call*-beëindiging onder een vastgestelde doelwaarde te houden. Dit leidt tot een toename van de blokkeringskans voor nieuwe *calls* toe, hetgeen een hogere ruimtelijke dichtheid van basisstations vereist. Dit hoofdstuk presenteert een analytisch model en evaluatie-methode, die voldoende eenvoudig zijn om een relatief snelle netwerkoptimalisatie toe te laten, doch realistisch genoeg om waardevol kwalitatief inzicht te verschaffen voor netwerk-planningsdoeleinden. In een uitgebreide numerieke studie wordt mobiliteit als een sleuteleigenschap geïdentificeerd, die niet genegeerd dient te worden in het radio netwerk-planningsproces. Daarnaast wordt gedemonstreerd dat de capaciteitsreservering ten behoeve van call handovers uitermate effectief is in het reduceren van zowel de kans op vroegtijdige *call*-beëindiging als de benodigde netwerkinvesteringskosten.

Hoofdstuk 9 concentreert zich op de prestatie-analyse van WLANs en presenteert een geïntegreerde modellering op pakket/*call*-niveau voor de evaluatie van *throughputs* en doorlooptijden in WLANs. Voor persistente data *calls* beschouwt het model op pakket-niveau het statistische gedrag van de individuele pakket-transmissies volgens het *carrier sense multiple access* protocol. Het model op *call*-niveau analyseert de dynamiek van het initiëren en afronden van de transmissie van data-bestanden, waarbij een *generalised processor sharing* discipline het WLAN grondbeginsel reflecteert om de beschikbare capaciteit op eerlijke wijze te verdelen over de opstaande data *calls*. Het geïntegreerde model is analytisch oplosbaar en leidt tot eenvoudige uitdrukkingen voor de verwachte *throughput* en doorlooptijd. Extensieve simulatie-experimenten bevestigen dat de modelbenadering uitermate nauwkeurig is voor alle beschouwde scenario's. Vervolgens wordt de gevalideerde evaluatie-methode gebruikt om de invloed van de verschillende WLAN toegangsmodi en de systeem- en verkeersparameters op de geleverde dienstkwaliteit te analyseren.

In een epiloog worden de belangrijkste behaalde onderzoeksresultaten samengevat en wordt een richting aangegeven voor relevant vervolgonderzoek.

ABOUT THE AUTHOR

REMCO Litjens was born on October 19, 1970 in Arcen. After attending the local primary school and enjoying his first math exercises at the age of five, Remco continued his education at the Collegium Marianum in Venlo, where his interest in the exact sciences led him to graduate with a Gymnasium β diploma in 1989. The same year he moved to Tilburg to study econometrics at the Katholieke Universiteit Brabant. The most memorable experiences during this period include the year-and-a-half he spent co-organising the successful congress 'Het ondernemende ziekenhuis' ('The entrepreneurial hospital') with four fellow econometrics students. In his final academic year in Tilburg, Remco spent a semester at Virginia Tech in Blacksburg, USA, taking classes in economics and operations research and enjoying American campus life. The subsequent semester proved pivotal in Remco's further life. Aided by his graduation professor Frank van der Duyn Schouten, a personal dream was fulfilled in joining professor Jean Walrand's research group at the department of Electrical Engineering and Computer Sciences (EECS) of the University of California at Berkeley, USA. It was during these six months at CAL, that Remco became fascinated by the world of mobile communications. With the defense of his graduation thesis 'Conflict-free scheduling of broadcasts in linear packet radio networks', Remco finalised his study in econometrics with a specialisation in operations research and was awarded his doctorandus diploma cum laude in 1994.

As his stay in Berkeley was appealing not only from an academic perspective, but even more so from a personal perspective, having met his future wife Mirjam, Remco enthusiastically grabbed the opportunity that professor Jean-Paul Linnartz of both CAL and the Technische Universiteit Delft offered him to participate in a joint project on automatic incident detection on intelligent freeways, carried out partially in Berkeley. Soon after the completion of this project, Remco entered the EECS graduate school at CAL with the original intention to pursue a doctorate. On the way this plan was altered and two semesters of research, taking and teaching classes later, Remco obtained his master of science degree with a perfect grade point average in May

1996 (thesis title: 'Conflict-free scheduling of broadcasts in wireless communication networks'). One principal reason for temporarily abandoning the Ph.D. objective, was the desire to travel around the globe. A few months after tossing his graduation hat at the Berkeley commencement, Mirjam and Remco set off on a 404-day trip through Africa, Asia, Australia, the South Pacific and North America.

Less than two months after those final troublesome days in Harlem, New York, Remco started his professional career as a technical scientific researcher at KPN Research, Leidschendam. Already in his job interview Remco had expressed a renewed desire to pursue a doctorate, and in the summer of 1998, he started his part-time Ph.D. research under the guidance of Richard Boucherie, first at the Universiteit van Amsterdam, later at the Universiteit Twente, Enschede. The 5 years that led up to the completion of this monograph were enriching and rewarding yet certainly arduous at times. Aside from the occasional acceptance notifications of submitted publications, definite highlights of this period were the conference, seminar and working visits to such diverse places as Barcelona, Bern, Cannes, Copenhagen, Dubrovnik, Helsinki, Karlskrona, Lillehammer, Lisbon, London, Osaka, Rhodes, Stockholm, and the memorable three month collaborations at NTT DoCoMo R&D Center, Japan. Arguably uncorrelated to any achievements presented in this monograph, on January 1, 2003, KPN decided to hand its research lab over to TNO, where it continued as a new institute under the name TNO Telecom.

After the defense of this monograph on September 12, 2003, Mirjam and Remco intend to take another five months off to travel through South America, after which a recharged Remco intends to pick up his work again at TNO Telecom and pursue new challenges.

SUBJECT INDEX