



# Performance Analysis of Broadband Wireless Networks

Péter Fazekas

Ph.D thesis

Scientific supervisors:

Prof. László Pap

Prof. Sándor Imre

Budapest University of Technology and Economics

Department of Telecommunications

September 2012

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Analytical model of mobile radio network</b>	<b>9</b>
2.1	Overview of mobile network model . . . . .	9
2.1.1	Cell capacity . . . . .	9
2.1.2	Arrival processes and customer classes . . . . .	11
2.2	Customer describing times . . . . .	13
2.3	User traffic model – general Markovian sources . . . . .	20
2.4	Base station model . . . . .	28
<b>3</b>	<b>Queueing model of the radio cell</b>	<b>31</b>
3.1	Service process . . . . .	31
3.2	The driving process . . . . .	36
3.3	Local balance and product form . . . . .	39
3.4	Approximate analysis of the finite capacity case . . . . .	43
3.4.1	Performance parameters . . . . .	47
3.5	Numerical results: Analysis of UMTS downlink channel . . . . .	48
3.5.1	Overview of basic UMTS radio operation . . . . .	48
3.5.2	Parameters used in the analysis and results . . . . .	51
<b>4</b>	<b>Calculating the residual session length distribution</b>	<b>60</b>
4.1	Modelling assumptions . . . . .	61
4.2	Calculating the Residual Session Length . . . . .	62
4.2.1	General Distributions . . . . .	62
4.2.2	Phase Type Models . . . . .	63

4.2.3	Determining handover probabilities . . . . .	65
4.3	Special scenarios . . . . .	67
4.3.1	Motorway model . . . . .	67
4.3.2	Homogeneous hexagonal layout . . . . .	68
4.4	Numerical results . . . . .	68
<b>5</b>	<b>Capacity and performance of 3G systems</b>	<b>72</b>
5.1	3G radio interface . . . . .	72
5.1.1	HSDPA operation . . . . .	74
5.2	Capacity analysis . . . . .	78
5.2.1	System description . . . . .	79
5.2.2	Capacity of Release '99 radio interface . . . . .	80
5.2.3	Capacity and performance of HSDPA . . . . .	87
5.3	Numerical evaluation . . . . .	90
5.3.1	Capacity of UMTS system without HSDPA service . . . . .	93
5.3.2	HSDPA capacity results . . . . .	98
<b>6</b>	<b>Conclusive remarks</b>	<b>101</b>
6.1	Future research . . . . .	103
	<b>Bibliography</b>	<b>106</b>
<b>A</b>	<b>Scientific publications of the author</b>	<b>118</b>
<b>B</b>	<b>New scientific results formulated in the dissertation</b>	<b>122</b>
<b>C</b>	<b>Theorems and proofs</b>	<b>128</b>
C.1	Proof of the existence of local balance equations (3.7) . . . . .	128
C.2	Alternative proof of the existence of equilibrium distribution 3.12 . . . . .	129
<b>D</b>	<b>List of Abbreviations</b>	<b>131</b>

## **Abstract**

This dissertation is devoted to the performance analysis of broadband cellular networks. The first part sets the general modelling framework of a cellular system. The model investigates session level performance parameters. The main contribution of this method is the incorporation of arbitrarily distributed session durations and cell residence times, along with the modelling of variable bitrate or bursty traffic sources. The resultant queueing model is analysed by the generalised version of the Kaufmann-Roberts formula. The second part investigates the effect of handover: a calculation method is shown, that enables the determination of the distribution of the time that remains from a communication session, when an already transferring customer hands over to an examined cell. The third part is focussing on the capacity issues of 3G cellular networks. A method is presented which enables the estimation of cell capacity under realistic circumstances. The main novelties are the incorporation of multipath propagation, multiple radio bearer types, user spatial distribution and the investigation of achievable useful capacity. HSDPA services and its interaction with Release'99 UMTS services is also analysed in terms of achievable HSDPA performance and cell throughput.

# Chapter 1

## Introduction

Following the mayor research areas within the broader field of telecommunications, we may conclude that that the topic of wireless multimedia networking deserves special attention today. Currently (July 2008) 30 out of around 140 EU financed research projects<sup>1</sup>, as part of the 7th research Frame Programs explicitly deal with wireless/radio communications, in topics spanning from physical device planning to mobile 3D TV broadcasting, having the total EU financial contribution of around 95 million Euros. Numerous other projects also – inevitably – consider wireless access, although the main profile is not on mobile communications.

The highlighted importance of wireless research, current and foreseen flowering of this area is based on the success of the related industry and the rapid deployment of cellular networks in the past 15 years. This is partially fueled by the exponentially growing capability of electronic devices. The the current IC technology is now approaching its maximum in terms of computing speed, but for the production of user friendly small portable terminals it is enough abundantly. This is supported by the ever improving quality of visualization and battery lifetime.

On the other hand, the enormous success of 2G digital cellular systems proves that customers are willing to use mobile devices. In parallel, we witnessed the spreading of personal computers and the supporting industries. Putting these into one small portable will assure customers' attention (if such a device is produced cheaply). In the past 2 years, after the deployment of HSDPA (High Speed Downlink Packet Access) services over 3G networks, we see the rapid spreading of the use of cellular Internet access.

Meanwhile the networking science also provided the extensions of the most popular protocols

---

<sup>1</sup>within the topic Pervasive and Trusted Network and Service Infrastructures of ICT (Information and Communication Technologies)

to support moving terminals. The mobile IP extension of today's basic Internet applications is available since 1996. The wireless ATM protocol had also been standardized during the end of 90's, but today WATM seems to be just an interesting historical idea. Currently no one really considers the deployment of end-to-end WATM services, ATM itself rather became a high-speed (but expensive) solution for carrying large amounts of aggregated user data traffic, rather than a transport mechanism of desktop.

But the clear requirement of provisioning QoS in wireless and wired networks forces engineers to introduce changes into the proven techniques, since the IP protocol family and its intended bearers were optimized for asynchronous, connectionless, distributed data communications. QoS is a must in future networks, since important applications (such as streaming audio and video, video telephony) require its information elements to be delivered with preserving timing relations and keeping the loss of information at a low level.

Keeping QoS parameters within a certain interval for each customer application is much more difficult in mobile environments, compared to wired networks, because of two main sets of reasons. One is the problems risen from using the error prone, low bandwidth, noisy radio channel. The other set is of the problems following the phenomenon that a user terminal may change its physical point of attachment to the network, while it continues transmission. This handover should happen seamlessly, with preserving the values of QoS measures. Additional delay and/or loss is not a surprise, because of necessary administrative messaging between the terminal and the network, or because of the lack of transmission capacity at the next network attachment point (usually referred to as access points or base stations).

The building and maintenance of a wireless network with such hard requirements requires modeling methods, that assist the performance evaluation of the system. As cellular network deployment is very expensive, the determination of the proper amount of necessary equipment is a must for mobile operators. Planning a cellular network is therefore a long and costly process. The procedure can be roughly divided into three main tasks:

- Dimensioning: this is the task where the number and placement of radio cells is determined. This phase can be viewed as rough capacity planning, as according to estimations on traffic volume and characteristics the required amount of radio resource is determined over an area, under the constraints of quality parameters.
- Radio planning: during this task the radio coverage is estimated, starting from the dimensioning plan, taking into account geographical properties, coverage requirements and other

constraints, such as possible placement of base stations. These two tasks are iteratively repeated, until both coverage and capacity requirements are fulfilled.

- Optimisation: according to field measurements and live network measurements, the coverage and cell capacity parameters are tuned.

The dimensioning task is where capacity and performance modelling solutions are needed. Such methods must be detailed enough to capture the main attributes of user traffic and network behaviour. Yet these models should be simple enough to get results soon. This dissertation is devoted to present such analytical methods, that can be utilized during cellular network dimensioning tasks.

Chapter 2 presents the basic general model and system description of a cellular system (including the modelling of customer behaviour), regardless the actual wireless technology used. The next part, Chapter 3 present the Markov model of the previously presented system, its analysis and some numerical results. The main novelty of this modelling is in the possibility to use general distributions – in place of the traditional exponential assumption – for describing customer behaviour (mobility, session duration), the presentation and inclusion of a new traffic modelling framework to model bursty or variable bitrate user generated traffic and the application of different traffic handling policies. Moreover as opposition to numerous papers, where infinite capacity is assumed and overload probability is given as quality measure, this approach considers finite capacity. The proper and exact interpretation of customer describing time variables is also often missing in the literature, this is shown and handled in the model proposed in this dissertation. This modelling can answer the question: what is the quality of the network, under given traffic load and user characteristics. Directly using it during planning will mean: what amount of radio capacity is needed over an area, in order to keep the quality parameters under a given threshold.

In the literature there are studies that use somewhat similar approaches that are presented in this dissertation. Since the basic and often referred work of Hong and Rappaport [1] a lot of effort was put into the research of queueing analysis of cellular networks. These works mainly focused on call level modelling and analysis of mobile telephone systems, later more general models with multiple traffic classes appeared. Usually continuous time queueing theory methods are used to evaluate network performance, some capacity sharing or admission control schemes, or to calculate several system parameters. The number of papers in the area is abundant, without the need of completeness a short review of some interesting ones follows.

In [2] a multidimensional Markov model of a radio cell is outlined. The authors supposed three types of traffic and the standard exponential time variables and evaluated call blocking and handoff failure probabilities in case of channel reservation for handover connections. The authors of [3] analyse blocking performance of systems consisting of multiple cells. They sketch the framework for proper model, then to obtain product form solution investigate special mobility cases (very slow mobility – the handoff rate tends to zero, very fast mobility – the handoff rate tends to infinity). At the end an approximate method of calculating handoff blocking is formulated based on the isolate evaluation of three cells. In [4] a network topology was considered consisting of microcells, that are overlaid by macrocells used to handle connections that cannot be served by microcells. Iterative algorithms are proposed to compute micro- and macrocell loads and then call incompleteness probabilities are derived. The work presented in [5] was devoted to analyse blocking performance in linearly placed cell arrays, with the dynamic assignment of a single channel for handover purposes. Closed form expressions are derived for handoff blocking and new call blocking probabilities. Multiservice wireless network with real-time and non-realtime connections was examined in [6]. In the paper the evaluation of different capacity sharing mechanisms was presented, where realtime traffic was allowed to occupy different amounts of capacity, according to the sharing method, and non-realtime traffic shared the remaining capacity. The performance of these sharing mechanisms was evaluated in terms of forced termination probability and the probability of having more than a threshold amount of capacity available to non-realtime traffic. In [7] an approach was provided for multimedia traffic, based on the performance parameter of cell overload probability. Here the connections were forced to occupy less capacity in case of lack of resources, this approach is similar to one policy what is presented in this dissertation. In the paper the good old exponential assumption was used for customer description. The work presented in [8] shows a connection level modelling framework for cellular systems. In this paper the exponential assumption of user describing time is released, also the required modelling modifications (residual lifetime of connections) because of non-exponential connection holding times and dwell times are described. This approach can be viewed as a subset of what is presented in this dissertation. The work presented in [9] is supposing voice calls (newly originated and handover traffic) and data calls. A finite buffer queueing model of a cell is set up, including queueing priority and guard channels for handoff calls. Closed form expression of the queue lengths is derived, as well as the Laplace-Stieltjes transform of the actual waiting time distributions. This study also incorporated the usual exponential assumption for channel holding time. In [10] two different handoff schemes were proposed and analysed,

containing guard channels preserved for handover connections. Preemptive and non-preemptive channel borrowing schemes were proposed; the analysis is based on a multi-dimensional Markov model of the system. In [11] a connection level Markov model was set up to analyse blocking and partial blocking schemes applied in OFDM (Orthogonal Frequency Division Multiplexing) systems. The framework uses either exponential, or hyper-Erlang distributions to describe connection lengths, the well established techniques for such analysis with fixed number of channels is here used to describe subcarrier allocation mechanisms. The approach of [12] was to set-up a closed queueing network model for a base station in OFDM based systems and use the given basic frameworks to analyse the effect of different frequency reuse schemes. The work presented in [13] considers cellular system with hybrid channel allocation scheme. This means that some channels are assigned to cells statically, but some channels are dynamically divided among several cells, taking inter-cell interference limits into account. Call blocking probabilities are derived, based on standard exponential channel holding time assumptions. The work presented in [14] is focusing on the queueing evaluation of a dynamic guard channel scheme in cellular networks. Again, customer describing times are supposed to have an exponential distribution. The authors of [15] stepped forward in terms of properly analysing a teletraffic framework in terms of handling cell dwell times (by means of Coxian distributions), however they stucked to fixed cell capacity and connections with exponential distribution, requiring unit capacity.

Later in this document more literature is shown, as the presentation of the modelling framework requires. During the elaboration of this framework, another interesting question had risen, namely how to determine the distribution of the (residual) duration of a communication session, that was initiated somewhere in the network earlier and arrives to the point of observation after some time has already elapsed. As mentioned, [8] partially touches this problem and determines the expected channel occupancy time in some special cases. The model presented in [4] also considers residual time variables. The authors of [16] investigate the effect of mobility on blocking performance, hence they provide means and Laplace transform of the holding time for connections initiated in a given cell, with exponential call holding time and general cell residence time distributions. Other papers often do not derive this quantity, either because the exponential assumption does not require this, or this descriptor is supposed to be given in the modelling. The approach shown in Chapter 4 of this dissertation takes network geometry and user mobility patterns into account and enables the use of general distributions.

The last topic investigated in this dissertation is presented in Chapter 5 and it is the problem of determining capacity of current 3G networks with HSDPA (High Speed Downlink Packet

Access) enabled. This topic deserved special attention in the past few years, and this is because the applied multiple access technology behind 3G radio networks. Namely, as CDMA (Code Division Multiple Access) is applied, due to interference reasons the offered capacity is not pre-determined, rather it itself depends on traffic and propagation issues, thus radio capacity, coverage and carried traffic are dependent on each other and are strongly coupled. The novelty of the approach presented in this dissertation is the method of handling the multiservice nature of UMTS, in contrast with the fact that most literature deals with capacity only in case of the presence of a single connection type. Moreover, this modelling captures the effect of multi-path propagation by means of the usage of distance-dependent orthogonality factor. The effect of users' spatial distribution and cell size is also incorporated and investigated, as well as the co-existence of traditional Release'99 UMTS traffic and HSDPA traffic on the same carrier frequency in a cell. Last, but not least, the fact that there are several HSDPA terminal types with different capabilities is also incorporated. The parameter under investigation is the average cell throughput, as this can be directly used for network dimensioning purposes. The analysis shown in this dissertation is focusing on the downlink, or forward link of 3G systems. Naturally, the motivation behind is that currently most 3G systems are deployed using FDD mode in paired spectrum, thus the same physical bandwidth is available for uplink and downlink traffic. However, the volume of downlink traffic is usually much higher than that of upload traffic, as consequence the capacity dimensioning task is usually performed for downlink.

The early works on CDMA system capacity issues [17][18] set the basic framework for CDMA analysis research. However, these works did not include multiple connection classes with heterogeneous signal to interference ratio requirements and transmission rates. Moreover, these works focused on the uplink performance, as they considered only symmetric voice calls and in this case the uplink is the bottleneck direction. As the 3G systems were standardized, from the beginning of 2000's several papers appeared targeting the WCDMA radio interface. Numerous papers deal with this problem using simulations, but we are rather interested in those that use analytical approach. The basic downlink pole equations were formulated and used for the estimation of used power in [19] and the approach presented here is the common base for a number of studies in the area. Results were shown for speech connections only, but the method of generalisation for multiple service classes was also outlined in the paper. The calculations shown in [20] are similar, but in this case authors consider the gain from soft-handover (macro diversity) as well. Again, total output power was calculated and presented numerically, but for single connection types only. In [21] a closed form expression was derived for the outage probability

and capacity was defined as bounded by the outage probability. The approach is based on an assumption, that was introduced by Viterbi [22] and appears quite often in the literature: as CDMA is having a soft capacity nature, the radio cell is modelled as an infinite server queueing system, resulting in the number of customers to have Poisson distribution. Then outage probability is defined as the probability of the total power exceeds the maximum available. This approach was used e.g. in [23]. Here the the outage in case of voice connections was determined using a Gaussian approximation of the total used power, in case of data connections lognormal approximation was used. In [24] the authors derived an analytical approximation method to examine the mean and variance of used power in UMTS downlink and modelled the total used power as having lognormal distribution. Based on this, in [25] they presented a Markov model for determining soft blocking and total blocking probabilities. In [26] an interesting approach was outlined, as the author defined a dimensionless quantity that serves as the amount of used radio capacity (which takes into account the multiple connection types and considers the random placement of users over the cell). As HSDPA services became standardized, papers appeared evaluating its performance. However, analytical approaches are less frequent, authors rather base their work on simulations. A flow level approach was shown in [27]. In this paper the basic methodology of evaluating HSDPA is well described, then two scheduling schemes are evaluated analytically and one more using simulations in different scenarios. In [28] all the necessary equations are written for joint UMTS/HSDPA performance evaluation, however the authors do not deduce results, but use the equations in computer simulations and present curves obtained by simulations. A clear work was presented in [29] in terms of the necessary equations. The effect of hybrid ARQ and the use of MIMO antenna systems was also included in the equations shown. Again, simulation results are available. Similarly, the authors of [30] present the necessary equations for UMTS/HSDPA analysis but evaluate these by means of simulations. The work presented in [31] derives a multidimensional Markov model of a 3G cell with HSDPA. This model considers uplink and downlink in parallel, taking into account fixed rate streaming connections and elastic data flows. The resulting quasi birth-death process is solved by matrix geometric methods. 3G specialties are taken into account by means of the derived maximum capacity in uplink and downlink. Interesting results are shown regarding the effect of uplink transfer rate on the achieved downlink performance. The technical report of Qualcomm [32] well summarises the capacity issues of 3G radio interface, however for Release '99 only circuit switched data (single service) is considered. Moreover, for HSDPA service only simulations are presented. The book written by the same operator [33] discovers very detailed aspects of 3G network capacity and

planning, however lacks the simple and compact formulation of multi-service 3G and HSDPA radio interface capacity.

The methods shown in this dissertation are analytical approaches to evaluate wireless network performance. The general analysis shown in Chapters 2 and 3 requires a fraction of time compared to simulation, because of the applied fast recursive solution. As the method presented is an approximate one, the accuracy is tested against simulations of the system as well. The results of Chapter 4 can be obtained using different approaches, as shown there. Again, as proof of concept, numerical results are compared to that of simulations. The evaluation of 3G systems presented in Chapter 5 relies on numerical computations based on presented analytical expressions. The results are compared with snapshot simulations in this Chapter as well.

# Chapter 2

## Analytical model of mobile radio network

In this Chapter I introduce and investigate a general analytical model of broadband cellular networks. The model is capable of investigating session level performance parameters of a single radio cell or sector. To obtain these performance measures the model requires very general and realistic assumptions regarding customer behaviour. Moreover the method is capable of handling flows with data rates that vary in time.

### 2.1 Overview of mobile network model

In this Section I present the overall modelling assumptions regarding the cellular network under consideration.

#### 2.1.1 Cell capacity

The cellular system considered during the elaboration of the analytical model presented in this document consists of a number of radio base stations. In practical cases base stations usually maintain more than one cells using sector antennas. In this case each sector is a cell, containing a number of radio channels allocated to the sector during radio network planning, according to the requirements of frequency reuse and radio capacity needed at a particular geographical location. Each cell can be characterised by the amount of transmission capacity it can provide for customers. However, in this document it is often supposed that a single base station maintains a single cell, thus the phrase cell or base station is used interchangeably. The model presented in this thesis does not require the assumption of any particular radio interface, nor does it deal with

the networking technology used to sustain communication among the base stations and between the cellular system and external networks. The base stations are characterized by the transmission capacity provided on their air interface, regardless the actual channel access and sharing method. This approach enables the investigation of wide range of current and future cellular networks.

However, the notion of capacity is not as simple and straightforward in a wireless environment, as in a wired network offering given link data rates. Namely, cell capacity is itself hard to define, as well as in general is not a time invariant constant. The reasons of these are the following:

- depending on their location, the perceived Signal to Interference and Noise Ratio (SINR) of different users is different, hence the achievable bitrate of customers (that is the cell capacity seen by a user if it were alone in the cell) is different
- due to multipath radio propagation and signal fading phenomena, the quality of the transmission channel is time-varying even for a stationary user, resulting in the variation of capacity for that user
- in modern mobile systems channel coding and modulation is chosen adaptively according to instantaneous channel state, resulting in the variation of the number of net information bits in a given time slot; that is the change of useful bitrate
- in today's systems radio resource has several capacity dimensions (e.g. in 3G systems time, channelization codes and transmission power are the capacity dimensions), that complicate the notion of capacity (e.g. transmission power cannot be translated into a transmission rate/capacity directly)

Despite these facts, calculations with a single, fixed capacity can be applied in the following cases:

- for some of the given radio resource dimensions of a system a fixed capacity may be directly assigned (e.g. the channelization codes in UMTS system, see the results of Section 3.5), independently of the channel conditions
- it is possible to define and serve data connections with given useful bitrates in cellular systems. In this case the applied modulation and coding allow for serving the customer with the given useful rate in case of bad channel conditions. For these given bitrates the necessary amount of physical resources is well defined (fixed) for at least some of the capacity

dimensions, hence capacity can be expressed in terms of multiples of such bitrates. As example, in UMTS system typically 64, 144 and 384 kbps useful bitrate connections are defined, the amount of channelization codes required for these is fixed.

- "best case" or "worst case" scenarios can be defined and evaluated, assuming a best or worst capacity case (e.g. worst case is when in all positions all customers may only use the most robust transmission format, hence the less bits can be transferred per time unit)
- taking all these into account, definition of and average or equivalent capacity can be calculated and the performance analysis of the systems is carried out with assuming this capacity. As example the work presented in Chapter 5 of this Thesis is on the definition of the average radio capacity in a multiservice 3G environment, but other authors (e.g. [26]) define the single dimensioned radio resource for 3G systems as well

### **2.1.2 Arrival processes and customer classes**

The model focuses on the performance of a cell. The incoming traffic of the examined radio access point is assumed to be known, independently of other base stations. The effect of the fact that customers are roaming throughout an area that is covered by several cells is taken into consideration by means of some of the user describing time variables, that is affected by other cells. When using the basic method presented here, one may calculate the performance of all the base stations in the network by applying this method for all base stations one by one throughout the cellular system.

The model is used to calculate session level performance parameters of a radio base station, namely the probability of blocking a communication session initiated within the cell, the probability of terminating a connection due to handover failure and channel utilization. These measures typically used to describe system quality when connection oriented services are used, such as conventional voice transmission, video telephony or video conferencing, streaming video and circuit switched data services. Using the notion of session rather than connection, this approach is capable of characterizing connectionless, packet switched services as well. In this context a session is a time interval when a customer is likely to generate data traffic pertaining to one information transfer session. For instance a session of web browsing is the time interval a user downloads and reads several pages or an FTP session is the time interval of transmitting one or more files. These sessions have definite beginning and termination, although during the

session transmission is completed by means of independent data packets or bursts. In the rest of this thesis the word session or connection is used interchangeably.

The model presumes that the number of customers that are not transmitting is very large in the coverage area of the cell under investigation and these users initiate communication sessions independently of each other and with small probabilities. Therefore the number of sessions initiated within the cell is approximated by a Poisson process. These connections are referred as new connections or new sessions, the arrival rate of such sessions is denoted by  $\lambda_N$ . Similarly, the number of active customers that roam in the vicinity of the cell is supposed to be large, and they initiate handover into the examined cell with small probability and independently of each other, therefore the incoming volume of active connections (handover) also follows a Poisson process, with the rate of  $\lambda_H$ . Obtaining these rates is completely out of the scope of this thesis, in practical cases the rates may be the results of measurements on the number of arriving connections, or can be calculated as it was suggested in [1].

While almost every paper dealing with the session level examination of cellular networks use the well tried Poissonian arrival processes, this may seem old-fashioned when investigating multimedia services. However, the assumptions of the previous paragraph are quite realistic, especially in densely populated areas with fairly large cells. In this case the assumptions inherently cause the incoming process of sessions to be Poissonian. Moreover, while numerous studies prove that *packet* arrivals are not Poissonian, according to the literature the inter arrival time of *connections* are still well described by exponential distribution. The authors of [34] and [35] showed that the arrival of TCP connections carrying user initiated Telnet and FTP sessions are well modeled by homogeneous Poisson processes within one hour intervals and in ten minutes intervals this is a good approximation for SMTP connections as well. In [36] the authors prove that the Poisson assumption for handover traffic is realistic. Nevertheless, some other models of incoming handover flow appear in the literature, in [37] a two moment representation of handoff traffic is used claiming that it is superior to the Poissonian assumption when a call generates numerous handovers. In [38] a two state MMPP is applied to model handoff traffic and the system is analyzed using this assumption, but with little indication of whether this approach is better or more realistic than the Poissonian. Similarly, in their recent paper [14] the arrival of new calls is assumed to be Poissonian, while the handover call arrival process is described by a two state MMPP. It also has to be noted that besides the cited papers above, the majority of the very rich literature dealing with analytical modelling of wireless networks assumes Poissonian arrival process for sessions.

I suppose that there are  $K$  different customer types and a mobile that arrives to the cell is of type  $k$  with probability  $\alpha_k$ ,  $k \in \{1 \dots K\}$ . These types are not identical with the possible service classes (for example QoS classes of UMTS). Users of different types may vary in terms of their generated traffic pattern and the length of their sessions (in this case they do belong to different traffic classes), or they might follow different mobility behaviour. The latter means that from modelling point of view customers of the same traffic class may belong to different user types if their mobility is different.

## 2.2 Customer describing times

In this Section the most important random time variables that describe users' mobility and duration of their transmission are described. Users of the system are characterized by two type of random time variables. The *session length*, or connection duration, or connection holding time is the time interval that lasts from the instant of initiating a connection until its termination. From the modelling point of view this is a continuous random variable. The general modelling allows the supposition that the holding time is different for different customer types, this could model the fact that session lifetime of a given service might depend on user mobility (e.g. web browsing sessions are generally lengthier in case of fixed terminals, than on fast moving terminals). However, useful and accurate data traffic models (including session lengths) that reveal such a dependence between mobility and traffic behaviour do not exist in the literature. The session length of a type  $k$  customer is denoted by  $\tau_L^k$ . The *dwel time* is a random variable characterizing the mobility of a user. It is often called cell residence time or cell sojourn time as well. This is the time interval that begins when the terminal enters the coverage area of a given cell (regardless it is transmitting or not) and lasts until the mobile leaves the cell. This time for a type  $k$  customer is denoted by  $\tau_D^k$ . We supposed that the mobility is independent of the communication pattern,  $\tau_L^k$  and  $\tau_D^k$  are independent. In the previous paragraph it was written that session duration may depend on mobility, but this would mean that the distribution of session length of a customer, that is described by an other distribution of dwel time will be different, than the distribution of session length with an other mobility behaviour. In the description, the two mobiles would belong to different classes. Hence it is not controversial with the assumption that the dwel time and connection lifetime is independent for a customer of a given class.

To create a queueing model of a radio base station we need the amount of time a customer occupies some capacity of the air interface. This time variable is frequently called *channel holding*

*time*. Generally this is not equal to either former time variables, rather it is calculated from those. We do not assume the session length nor the dwell time to have exponential distribution (although this is a very common supposition in the literature). Considering this, the session length and the dwell time do not have memoryless distribution, thus a refinement of the interpretation of session length and dwell time is necessary before the expression of the channel holding time.

Regarding the duration of a mobile's connection, the model requires the amount of time that lasts from the instant the connection attaches to the base station, until the instant of connection termination. For sessions initiated inside the examined cell this time is equal to the session length (the connection attaches to the base station at the instant of its initiation). Handover connections are set up somewhere else in the network, thus some time elapses until the mobile hands over to the examined cell. We are interested in the amount of time that remains from the instant of handover until connection termination. We refer to this time as *residual session length* and denote it by  $\tau_{L,R}^k$ . Considering exponentially distributed session lengths,  $\tau_{L,R}^k$  would have the same distribution as  $\tau_L^k$  due to the memoryless property of this distribution. The model presented here allows the use of more general distributions, in this case the residual session length has different distribution. Chapter 4 of this dissertation is dedicated to the calculation of  $\tau_{L,R}^k$  using the information on network topology, the session length and the dwell times. At this point it is enough to suppose that the residual session length is available somehow when analysing a radio cell.

The notion of dwell time also requires differentiation between the users of handover and new connections. For a customer that arrives to the cell after a handover, the dwell time is defined as described above. Regarding new connections, the notion of *residual dwell time* is introduced, denoted by  $\tau_{D,R}^k$ . This time interval begins at the instant of initiating the session in the cell and finishes when the mobile leaves the coverage of the base station (regardless it is still transmitting or not).

Assuming exponentially distributed dwell times is very common in the literature, this would make the notion of residual dwell time unnecessary. Other authors ([1][39][40]) derive the dwell time separately for handover and new connections. Another method is applicable to determine the distribution of the residual dwell time if we suppose that the system consists of homogeneous cells. This homogeneity means that the dwell time distribution is identical in all the cells of the system. A user that is constantly roaming throughout the area is supposed to initiate a session at a time instant that is evenly distributed along a very long time period. Then the problem of the residual dwell time is identical to the problem of travelling hippie presented in Kleinrock's

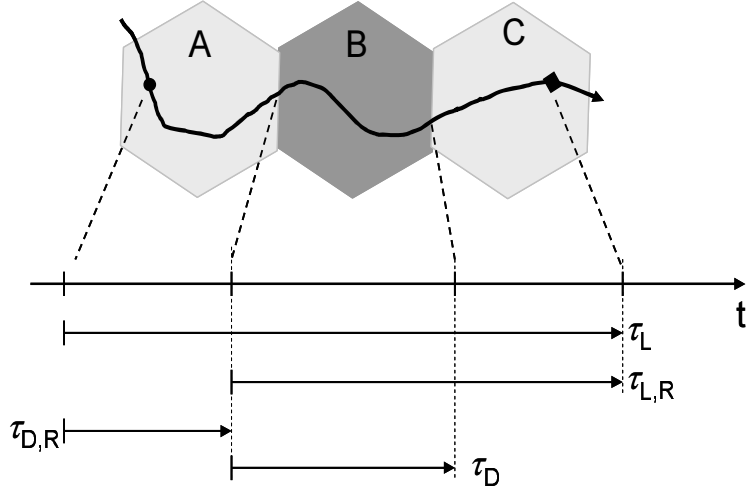


Figure 2.1: User describing times

famous book [41].

As the final result for the residual dwell time distribution is somewhat not intuitive, it is worth presenting its derivation shortly. For this, we divide the problem into two parts. First we suppose that we look at a related problem, where in every dwell time interval a connection would be initiated. So the question is in this case, that what is the distribution of the residual time that lasts from the connections initiation until the dwell time's end, given the probability density function and cumulative distribution function of the dwell time of a type  $k$  customer is denoted by  $f_D^k(t)$  and  $F_D^k(t)$  and the connection is initiated evenly over that given dwell time interval. According to this evenly distributed assumption, if the dwell time were actually of length  $x$ , both the initiation instant and the residual dwell time would have even distribution with density function  $\frac{1}{x}$ . Hence the joint distribution of the dwell time length and that of the residual dwell time could be expressed as

$$Pr(y \leq \tau_{D,R}^k \leq y + dy, x \leq \tau_D^k \leq x + dx) = \frac{1}{x} dy \cdot f_D^k(x) dx, \quad (2.1)$$

From here, we would get the pdf of the residual dwell time after integrating the above expression over  $x$ , from  $y$  to infinity

$$\tilde{f}_{D,R}^k(y) = \int_y^\infty \frac{1}{x} f_D^k(x) dx, \quad (2.2)$$

where the notion  $\tilde{f}_{D,R}^k(y)$  in (2.2) is to indicate that this is not the residual dwell time pdf what we were looking for, rather the one that would result in the special case when in every dwell time intervals a connection would be initiated.

To arrive to the final result, we should incorporate the effect of length biased sampling. Namely this is the fact that in our problem it is more likely that the user initiates call in a particular dwell time interval that is longer (it is more likely to pick a longer interval than shorter ones, as we supposed that the instant of picking is evenly distributed over a very long period). This eventually means that the pdf of the dwell times that will be actually "chosen" to initiate a call within is *not*  $f_D^k(x)$ . This means that although the form of (2.2) is appropriate, it is not the dwell time's pdf, but the *chosen* dwell time's pdf should appear in (2.2). Let us denote this latter pdf by  $\hat{f}_D^k(x)$ . With this assumption, we can write

$$\hat{f}_D^k(x)dx = K \cdot x \cdot f_D^k(x)dx, \quad (2.3)$$

where the left hand side is the probability of having the "chosen" interval to be "around"  $x$  and the right hand side expresses that this is proportional to the interval length  $x$ . The value of factor  $K$  appears when we integrate (2.3), as this should result in one. From this, we get

$$K = \frac{1}{\int_0^\infty x \cdot f_D^k(x)dx}, \quad (2.4)$$

that is the reciprocal of the mean dwell time! Substituting now this  $\hat{f}_D^k(x)$  into (2.2) instead of  $f_D^k(x)$ , we arrive to the pdf of the residual dwell time distribution, namely

$$f_{D,R}^k(y) = \frac{\int_y^\infty f_D^k(x)dx}{\mathbf{E}[\tau_D^k]} = \frac{1 - F_D^k(y)}{\mathbf{E}[\tau_D^k]}, \quad (2.5)$$

where  $\mathbf{E}[\tau_D^k]$  denotes the expected value of  $\tau_D^k$ .

Figure 2.1 illustrates the proposed time variables as a mobile follows a route through the cellular area. The black circle denotes the point of session initiation, the residual dwell time in cell A is depicted in the Figure. From the instant of handover into the examined cell B begins the residual session length, as well as the dwell time for cell B. As we can see, the connection is terminated in another cell, denoted by a black square.

Considering the above interpretation of residual session length and residual dwell time, the channel holding time is expressed as

$$\tau_{CH}^{k,N} = \min(\tau_{D,R}^k, \tau_L^k), \quad \tau_{CH}^{k,H} = \min(\tau_D^k, \tau_{L,R}^k). \quad (2.6)$$

for new and handover connections respectively. From now on in this dissertation the superscript H denotes a variable describing handover connections, N refers to newly initiated sessions within the cell, R denotes the residual value of a variable. Expression (2.6) simply means that a user

may terminate its occupancy of the channel either by means of finishing the connection, or by means of handing out of the cell. Thus either the (residual) dwell time or the (residual) session length is shorter for a customer, this will be the channel occupancy time.

In this dissertation I assume that all the customer describing times are modelled by having phase type distributions. From this point sometimes I use the short notion PH instead of writing phase type. PH distributions were first introduced by Marcel Neuts [42] and are composed as a mixture of a number of exponentially distributed phases. An initial probability vector determines the first phase, than upon ending a phase the next phase or the termination of the process is chosen according to a transition probability matrix. In other terms a PH distributed time is the time a finite state continuous time Markov chain reaches an absorbing state. Using this latter interpretation, the phase type distribution is characterised by the initial probability vector of the Markov chain and its infinitesimal generator matrix. Supposing that the states of the Markov chain are numbered such that the absorbing state's number is the biggest, the infinitesimal generator matrix of the chain has the following structure:

$$\begin{bmatrix} \mathbf{T} & \underline{T}^0 \\ 0 \dots 0 & 0 \end{bmatrix},$$

where  $\mathbf{T}$  contains the rates among non-absorbing states and  $\underline{T}^0$  is the column vector of rates from each state to the absorbing state. The sum of the elements of a row of the infinitesimal generator matrix must be equal to 0, therefore  $\underline{T}^0$  is determined by  $\mathbf{T}$ , namely  $\underline{T}^0 = -\mathbf{T} \cdot \underline{h}$ , where  $\underline{h}$  is a column vector containing 1s. Thus the distribution is well described by the initial probability vector  $\underline{t}$  and matrix  $\mathbf{T}$ . For this reason, in the following discussions of this dissertation I also use the general notion of  $\text{PH}(\underline{t}, \mathbf{T})$  to refer to a phase type distribution with these descriptors. The cumulative distribution function and the probability density function of a phase type distribution with the above parameters has the form of:

$$F(x) = 1 - \underline{t}e^{\mathbf{T}x}\underline{h}, \quad f(x) = \underline{t}e^{\mathbf{T}x}\underline{T}^0. \quad (2.7)$$

Numerous distributions are known and used widely that are PH distributions with special phase structure. The exponential distribution itself is the simplest PH, a bit more complex PHs are the Erlang, the sum of exponentials (which is a generalisation of Erlang by letting different means of each exponential phase) and the hyperexponential distribution. The acyclic PH distributions contain a linear sequence of phases, but the process may terminate in each phase with a certain probability, or continue with the next phase. This distribution is also used quite frequently and it is often referred as Coxian distribution. The sum of hyperexponentials (SOHYP)

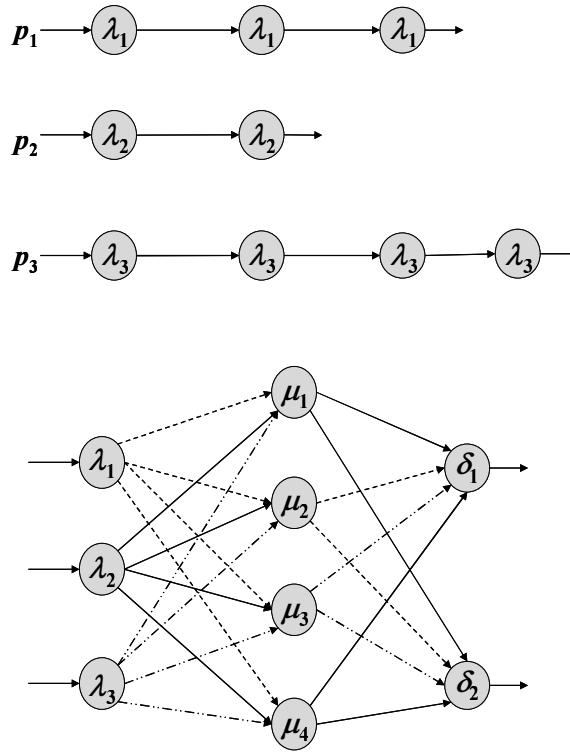


Figure 2.2: Hyper-Erlang (upper) and SOHYP distributions

and the hyper-Erlang distributions have even more complicated phase structure with generally more phases than the previous ones. The hyper-Erlang is constructed as choosing an Erlang with certain probability from a group of Erlangs, each having distinct phase numbers and phase means. The SOHYP is a concatenation of different hyperexponential distributions, with different means of distinct phases and with same set of branching probabilities from each phase of one "layer". On top of Figure 2.2 an example of the hyper-Erlang distribution is shown, the bottom of the picture depicts a SOHYP distribution. The different dashed lines of the latter are simply for better visibility purposes.

Studies show ([43][44][45][46][47][48][49]) that most distributions, even heavy tailed ones can be properly and effectively approximated by an appropriately chosen phase type distribution. Moreover, if statistics are available on a random variable with unknown distribution, a phase type distribution can be chosen that follows the statistics of the variable. Given these reasons it is quite general and yet realistic assumption to model all the user describing times as having a phase type distribution, since this assumption may include models with other proposed distributions, after the fitting of an appropriate PH, or this model can be used when real measurement data is

available about user describing times.

Thus the session holding time of type  $k$  new connections is phase type distributed  $\text{PH}(\underline{l}^{(k)}, \mathbf{L}^{(k)})$ , the residual session length of handover connections is described by  $\text{PH}(\underline{l}^{(R,k)}, \mathbf{L}^{(R,k)})$ . The dwell time of handover connections and the residual dwell time of new sessions have phase type distributions  $\text{PH}(\underline{d}^{(k)}, \mathbf{D}^{(k)})$  and  $\text{PH}(\underline{d}^{(R,k)}, \mathbf{D}^{(R,k)})$  respectively. In Section 3.1 we show that in this case the channel holding time is also phase type distributed.

This distribution family also sometimes appeared in the literature as model of some customer describing time variable. In [50] the author used sum of exponentials as model of the dwell time, later he modelled this time to follow SOHYP distribution in [51], but the connection holding time was still exponential. The authors went a bit further in [52] and analyzed cellular base stations with the session length also having SOHYP distribution. In [53] the dwell time was modelled by hyper-Erlang distribution, the same topic and arguments regarding the validity of this model was abundantly covered by one of the previous article's authors in [54] and [55]. The same distribution was used in [56] but as model of session length. The authors of [8] derived formulas supposing Erlang distributed call holding times and general dwell times only known by its mean, the effect of this assumption on the call completion probability was examined in [57]. General PH distribution modelled the channel occupancy time in [58] and the authors calculated system parameters by solving the resultant M/PH/ $n$  queue. In [59] both the session length and dwell time was modelled by general PH distributions. The time a user resides within the overlap area of two cells was analyzed in [60], that has significant relevance when investigating soft handover. The authors fitted numerous distributions onto numerical data and found that the hyper-Erlang distribution is the best approximation. The authors of [11] also analysed the exponential, Erlang and hyper-Erlang distributions as session length distributions in their work. The work shown in [15] elaborated teletraffic modelling framework with Coxian approximation of the dwell time also.

Other distributions (different from the conventional exponential assumption) also frequently appear in the literature regarding user describing times. To show some insights of such efforts a number of references given here. After fitting an appropriate PH distribution to any of these, our general model covers all these cases. Regarding the dwell times, in probably the most often referred early article in the field of teletraffic modeling of cellular networks ([1]) the authors derived special density functions of the dwell times of new and handover customers, depending on the cell radius and speed of mobiles. Based on this work the authors of [39] and [40] introduced

the generalized gamma distribution as model of the dwell time, with pdf

$$f(t) = \frac{c}{b^{a \cdot c} \Gamma(a)} t^{a \cdot c - 1} e^{-(t/b)^c}, \quad (2.8)$$

where  $a$ ,  $b$  and  $c$  are parameters dependent on the cell radius and on the fact that the call is a handover or new connection,  $\Gamma(a)$  is the gamma function defined as  $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$ . A slightly different model, gamma distribution was used in [61] and [62] to describe cell residence time, while in [63] lognormally distributed dwell times were used. In [64] authors claimed that the dwell time is well approximated by Pareto distribution. Interesting result was shown in [65], where the session length of network game applications was described by Weibull distribution based on experimental data (although in [66] the good old exponential distribution seemed appropriate to describe game lengths, also based on measurements). The work presented in [67] and [68] investigated the holding time in PAMR (Public Access Mobile Radio) systems, based on measurement data. The authors found that hyper-Erlang and mixture of lognormal distributions are the best approximations. The authors continued measurements on the cellular network of Barcelona and found that the channel holding time in this system is also well described by the mixture of 3 lognormal distributions ([69],[70]). The authors concluded their previous results in [71]. The authors of [72] also presented their work with the assumption of gamma distributed dwell times.

After reviewing a number of previous customer describing time models we may conclude that assuming these to follow some general PH distribution is a realistic proposition, covering the most of the models presented in the literature.

## 2.3 User traffic model – general Markovian sources

In this Section the model describing customers' generated traffic pattern is described. When modeling cellular networks with other applications besides speech connections – such as web browsing, streaming video viewing, on-line game playing, video telephony, etc. – one has to take into account that these sources do not generate data at constant rates, rather aperiodic bursts and silences follow each other, or data rate varies in time during a session. Although numerous studies were published about analysis of wireless networks with different but constant transmission rate connections (see references in the Introduction), it is very difficult to efficiently demonstrate the validity of these approaches in the multi-service wireless networks of today and the near future.

One argument could be the assumption of circuit switched data connections with different rates, that carry the workload of various applications, similar to the HSCSD (High Speed Circuit Switched Data) service on GSM networks. Although this interpretation is a step closer to the concept of multimedia wireless systems it is unrealistic in sight of the operation of packet switched services in current and future networks. Moreover it would be very inefficient and capacity wasteful to maintain a fix rate connection at a bursty source's disposal. Another approach would be to substitute each source with a virtual CBR (Constant BitRate) one with rate equal to the source's effective bandwidth. Effective bandwidth concepts were introduced in the early 90's to somehow describe bursty sources by a single transmission rate, yet keeping the ability of examining system behavior, such as queueing delays, connection admission control, etc. ([73], [74],[75],[76]).

At first sight it might not seem straightforward to ensure the variability of user data rate on the radio interface. Regarding random access wireless networks, such as the IEEE 802.11 family the medium access method inherently causes the data flow to be bursty: the terminals compete for the channel wherever a bulk of data is ready to be transmitted, the winner transmits using the total capacity of the channel, then remains silent until another transaction. Actually it is not the realization of bursty transmission, but communication using time variable data rates that needs some comments. Considering networks with organised medium access, time division, code division and OFDMA systems dispose current and future significance. In TDMA networks user rate variability is assured by means of variable amount of time slots allocated to a terminal in different time frames. Usually the customer notifies the central infrastructure about its capacity requirement in the following frames (using either explicit signalling channel or piggybacking this information attached to application data sent earlier) and a scheduler decides the number of time slots allocated to customers.

To fulfill user rate variability in CDMA networks is a bit more tricky (burstiness is again straightforward: the customer simply transmits or not with the code allocated to him). Although numerous papers deal with multi-service CDMA networks, this usually means the assumption of different, but constant rates per user class. However the basic ideas of assuring different data rates can be used to outline the way of providing time varying data rate for a customer. One approach is to use different spreading gains – utilizing orthogonal variable spreading factor (OVSF) codes –, for different bitrates (e.g. [77],[78],[79]). In this case bits of higher rates are spread by shorter codes, therefore by maintaining the same chip rate more bits are transmitted during a time interval than with longer codes. The other approach is to use multicode CDMA

transmission<sup>1</sup> (see e.g. [80], [81], [82]). In this case user data is transmitted as parallel low rate flows, each spread by unique codes, thus creating parallel sub-channels via the different codes. Varying user rates are then achieved by the varying number of parallel streams. To assure variable rates during a communication, the former approach would require to change the spreading factor (thus the spreading code) during a connection according to the actual required bit rate. This demands some special signalling to inform the receiver about the new code it has to despread the user signal with, or some complex blind rate detection algorithm has to be implemented in the receiver side. In the case of applying multicode CDMA this additional signalling is not necessary, since the receiver only has to despread all the parallel flows constantly, gaining no data from those sub-channels that are not used for transmission in case of lower bit rates. But if the system is planned for the maximum traffic carrying capacity, the codes not used by a source for a given period may be allocated to other sources. In this case again special signalling is required to inform the receiver about the instantaneous set of codes the source is transmitting with. Moreover, the so-called *code blocking* phenomenon (see references above and later Section 3.5.1 in this thesis) also complicates the use of OVSF codes or multiple codes for variable bit rate communications.

To capture the variable and bursty nature of general user traffic, in this dissertation I suppose Markovian traffic sources. This means that the data traffic generated by a connection is characterised by a finite state continuous time Markov chain. Each state of this chain is assigned with a transmission rate (that might be zero as well), meaning that the customer is able to transmit with a finite set of possible data rates. The customer starts its transmission with a rate that is determined by the initial probability vector of the underlying Markov chain and it keeps sending data with this rate for an exponentially distributed time, then the underlying chain jumps into another state. A state transition may result in the change of transmission rate (if the new state is assigned with a different rate), but the data flow may continue with the same rate as well, since more states may be assigned by equal rates. The traffic pattern of this model is characterized by the infinitesimal generator  $Q^{(k)}$  for a type  $k$  connection, the first rate upon session initiation is determined by the initial probability vector  $\underline{q}^{(k,N)}$ . The transition rates of the traffic describing Markov chain are obviously the same for new and handover customers but the initial probability vectors are different. It is because handover sessions were set up earlier than attaching to the examined base station, therefore the underlying Markov chain jumped several times until the instant of handover. We suppose that the change of transmission rates is very fast compared to

---

<sup>1</sup>often abbreviated MC CDMA which can be mistaken with multi-carrier CDMA

the dwell times or the session lengths, thus when a handover connection attaches to the base station it has been communicating long enough that the traffic describing Markov chain reached its equilibrium. Therefore the initial probability vector of handover connections  $\underline{q}^{(k,H)}$  is supposed to be the steady state distribution of the chain, calculated by solving the well known

$$0 = \underline{q}^{(k,H)} \cdot \mathbf{Q}^{(k)}, \quad \underline{q}^{(k,H)} \cdot \underline{h} = 1 \quad (2.9)$$

system of equations, where  $\underline{h}$  denotes a column vector with 1s in its each position.

This type of traffic model is quite common in the literature. Such a Markovian model was used in [83] to describe video sources with single or two activity factors as well as the aggregate traffic of several video sources. This type of source model was supposed in [84] to describe bursty video traffic, in [76] this general model was used to calculate the sources' effective bandwidth. In [85] a superposition of multiple on-off sources was used to model VBR traffic, that is also a case of the general Markovian model. Similar approach was applied in [86] and [87] to describe a multimedia source as the superposition of on-off monomedia traffics. The on-off model itself that appears frequently in the literature as description of speech transmission with voice activity detection or as model of other services is the simplest form of the proposed Markovian traffic model.

This approach of multi-rate Markovian traffic sources along with the ability of phase type distributions to follow almost any arbitrary distribution allow us to introduce very general models of bursty sources, yet exploiting the Markovian property of individual phases and thus allowing the use of well-established queueing theory methods to examine systems with such sources. One straightforward model is to use on-off sources with generally distributed on and off period durations. To indicate the significance of this approach we refer [88], where the authors investigated the capabilities of on-off sources with arbitrary distributions to model traffic generated by ATM endpoints. According to measurement data, a web traffic model with Pareto distributed off periods and on periods dependent on web page size distributions was used in [89] and in [90]. Similar model but with lognormally distributed think times between web pages was claimed in [91]. The authors of [92] also found the on-off model to appropriately describe user web traffic, but with Weibull distributed on and off times. Sources with heavy-tailed on periods were analysed in [93] also. The effect of heavy tailed on periods was simulated in a P-persistent CSMA medium access environment in the recent paper [94].

In order to reproduce these general on-off models as Markovian sources, appropriate PHs should be fitted to the distributions of the on and off durations and the Markovian model is the

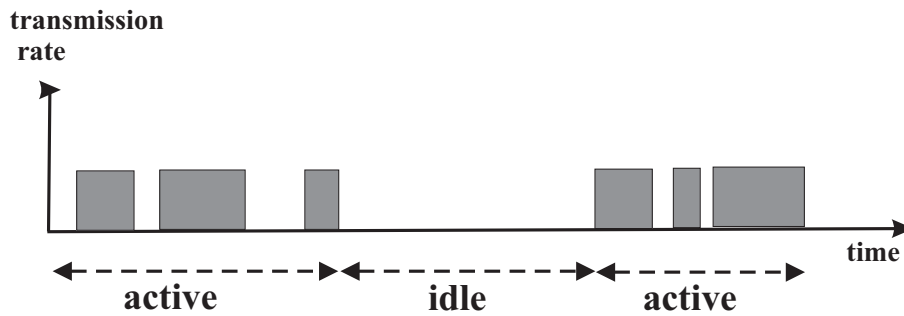


Figure 2.3: Traffic pattern of general bursty source

result of connecting the phases of the two PHs with finishing rates of the first PH multiplied by the initial probabilities of the second PH. This approach is described fully in the subsequent paragraphs as part of a more general traffic model introduced in this thesis.

Data traffic of packet-switched applications is often generated during active periods with idle time intervals between them. It is common that during active periods the transmission is not continuous in fact, rather consist of periods of sending data packets or files (bursts) interrupted by short silent intervals (think times). Web browsing is a typical application that is well characterized by this type of source model: usually several files are downloaded with short think periods until the customer finds the information he looks for, then follows a longer idle interval while the user studies the desired page. This type of model was suggested in [95] to describe traffic generated by interactive applications, such as web browsing, email and query/response information services (although numerical results were presented supposing constantly active customers with Pareto distributed think times). Similarly active periods with on-off intervals and inactive off periods modelled web user traffic in [96] emphasizing that the inactive off periods should not be ignored to get a realistic model. A good overview of earlier source traffic models for wireless networks is summarised in [97], the models listed here are basically in alignment with what was written earlier. Figure 2.3 shows an example of the traffic pattern generated according to the proposed source model. Note that this description with idle periods, silent periods and burst may be used as model with finer granularity: in this case a burst would be the transmission of a packet, silent period would be the time a packet is generated and the idle period would be transmission gap between packet bursts. To describe such a source as Markovian, we suppose that the active interval's length is a continuous random variable with pdf  $a(t)$ , the idle periods' distribution is described by the density function  $i(t)$ , the length of the bursts during the active period is de-

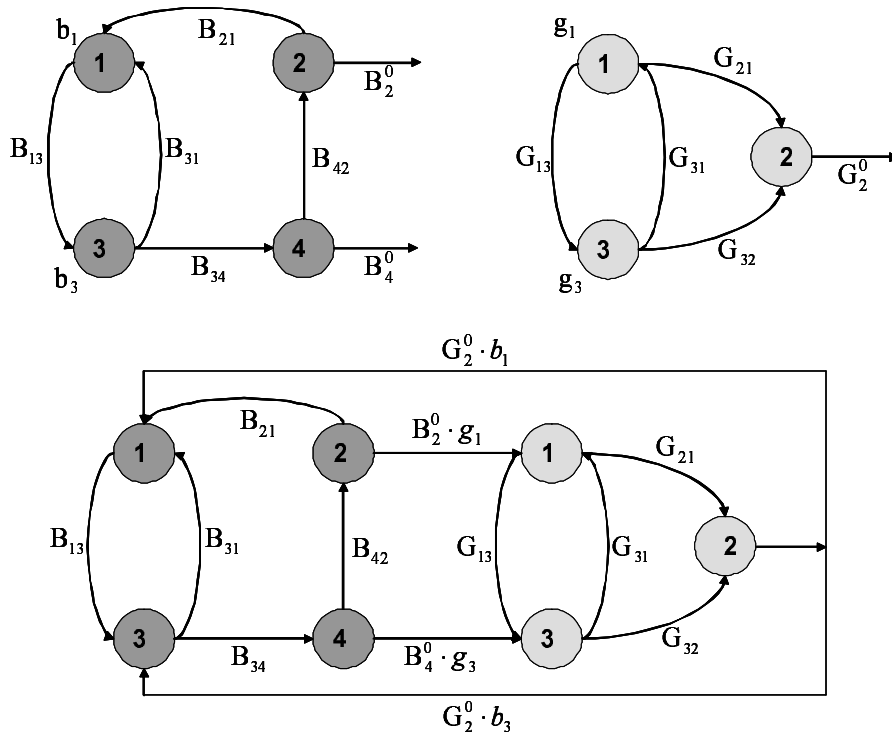


Figure 2.4: Markov model of active periods

scribed by the density  $b(t)$ , that of the short silent periods between bursts is denoted by  $g(t)$ . Let us assume that each distribution is approximated by a properly fitted phase type distribution, or only samples of the above time variables are available, and phase type distributions are chosen to fit the experimental data. The parameters of the four phase type distributions are denoted by  $(\underline{a}, \mathbf{A})$ ,  $(\underline{i}, \mathbf{I})$ ,  $(\underline{b}, \mathbf{B})$  and  $(\underline{g}, \mathbf{G})$  respectively.

First we create the model of active periods, with bursts and short silent periods between them. If a general ON-OFF model is needed (without the longer active and idle periods), with arbitrarily distributed ON and OFF times, this approach should be used to obtain its description. To achieve this, the phases of the silent period distribution are connected to those of the burst length distribution. The rates between the two groups of phases are the following: those phases that correspond to the non-zero elements of the initial probability vector  $\underline{g}$  are connected to those phases of the burst length distribution that correspond to the non-zero elements of the finishing rate vector  $\underline{B}^0$ . The rate between phase  $k$  of the burst length distributions and phase  $l$  of the silence distributions is  $B_k^0 \cdot g_l$ . To model the repetitive nature of bursts and silent periods those phases of the latter's distribution that correspond to the nonzero elements of its finishing rate

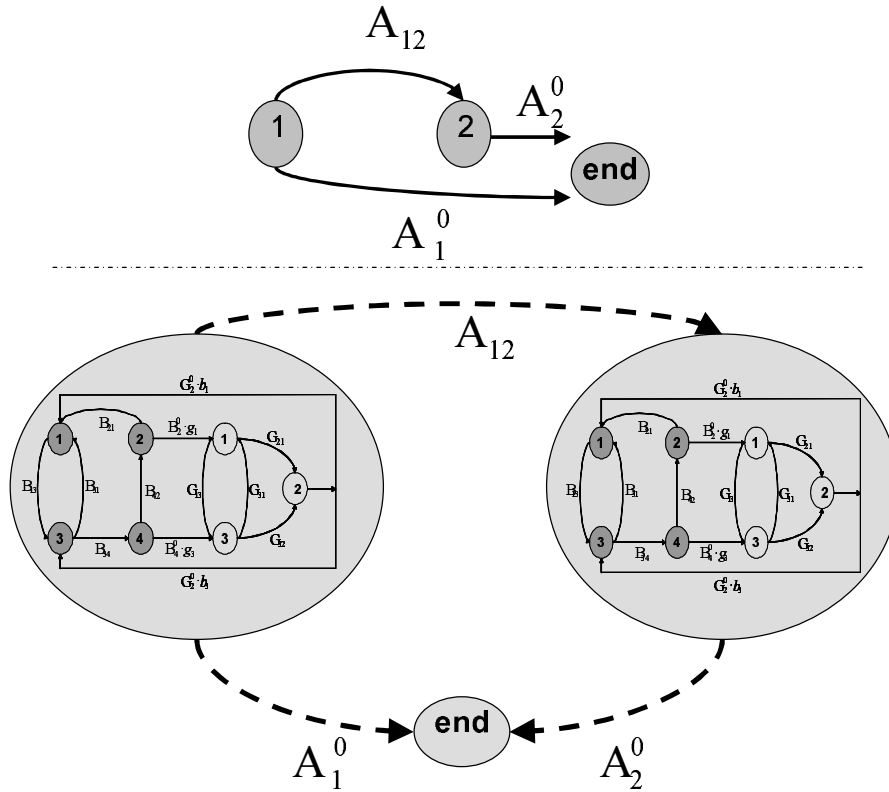


Figure 2.5: Markov model of active periods, including duration

vector are fed back to the phases of the burst period, with rates  $G_k^0 \cdot b_l$  between phase  $k$  and  $l$  of the two distributions. Figure 2.4 shows an example of creating the model of active periods. The upper left and right graphs show the phase type model of the burst and silent periods respectively. Below is the joint Markov model with the proposed connections among the two groups of states. It is easy to see that using appropriate numbering of states, this aggregate active period model has its transition rate matrix and initial probability vector with the following form:

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{B} & \underline{B}^0 \cdot \underline{g} \\ \underline{G}^0 \cdot \underline{b} & \mathbf{G} \end{bmatrix}, \quad \underline{a}^* = [\underline{b}, \underline{g}]. \quad (2.10)$$

Here the parameters are denoted by  $\mathbf{A}^*$  and  $\underline{a}^*$  since this model does not capture the duration of the active period (described by  $(\underline{a}, \mathbf{A})$ ), but the durations of bursts and silences within the active term.

The next step of creating the general bursty source model is to achieve that this active period lasts for a phase type distributed time with parameters  $(\underline{a}, \mathbf{A})$ . Thus the previously described model of the active period is taken as many times as many phases the length distribution of the

active period has. Each phase of each group is connected according to the rates of the active period distribution, namely phase  $j$  of group  $k$  is connected to phase  $j$  of group  $l$  with rate  $A_{kl}$ . Figure 2.5 shows an example of creating the model of active periods including its duration. In this example the length of active periods is supposed to follow a two phase acyclic phase type (often referred as Coxian as well) distribution, represented at the top of the Figure. The Markovian model of Figure 2.4 is used to describe bursts and silences within the active period. The thick dashed connectors of the figure represent that each state of the two groups are connected with appropriate rates. It is easy to see, that when the states of this latter model are enumerated appropriately, the resultant phase type model of the active period (including the alternation of bursts and silent periods) has the following descriptors:

$$\mathbf{A}_{\text{eff}} = \mathbf{A} \oplus \mathbf{A}^*, \quad \underline{a}_{\text{eff}} = \underline{a} \otimes \underline{a}^*, \quad (2.11)$$

where  $\otimes$  and  $\oplus$  denote the Kronecker product and Kronecker sum respectively. The Kronecker product of matrix  $\mathbf{A}$  of size  $m \times n$  and  $\mathbf{B}$  of size  $k \times l$  is a matrix of size  $mk \times nl$  defined as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} A_{11} \cdot \mathbf{B} & A_{12} \cdot \mathbf{B} & \dots & A_{1n} \cdot \mathbf{B} \\ & \dots & \dots & \\ A_{m1} \cdot \mathbf{B} & \dots & \dots & A_{mn} \cdot \mathbf{B} \end{bmatrix} \quad (2.12)$$

The Kronecker sum is defined on quadratic matrices  $\mathbf{A}$  of size  $m \times m$  and  $\mathbf{B}$  of size  $n \times n$  is a matrix of size  $mn \times mn$  calculated as:

$$\mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}, \quad (2.13)$$

where  $\mathbf{I}_n$  and  $\mathbf{I}_m$  denote the identity matrices of size  $n \times n$  and  $m \times m$  respectively.

Regarding an equivalent problem that rises during the elaboration of the service time distribution of the presented queueing model, it is shown in Section 3.1, Theorem 3.1.2 that this distribution with parameters in (2.11) has the same density function as the one with descriptors  $(\underline{a}, \mathbf{A})$ , thus this model appropriately describes the duration of the active period as well.

Finally, to include idle periods in the source model, similar approach is needed that was used when creating the active period model. The phases of the idle period distribution are connected to those of the total active period distribution and its finishing phases are fed back to the active period distribution with finishing rates multiplied by the initial probabilities of the other distribution. Thus, the infinitesimal generator and the initial probability vector of the final source model is constructed as:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_{\text{eff}} & \underline{A}_{\text{eff}}^0 \cdot \underline{i} \\ \underline{I}^0 \cdot \underline{a}_{\text{eff}} & \mathbf{I} \end{bmatrix}, \quad \underline{q} = [\underline{a}_{\text{eff}}, \underline{i}]. \quad (2.14)$$

This source model was shown as a general ON-OFF model, but it is straightforward to suppose not only 2, but arbitrary number of possible transmission rates in the model, let us denote this number now by  $R$ . More specifically, different possible rates could be given, with all the PH distributions of the duration of a burst with the given transmission rate. Compared to the previous case, the difference is that one more parameter is needed in this case, namely the probability, that after finishing the burst with a given rate, the transmission continues with another one. Let us suppose that these probabilities are collected in matrix  $\mathbf{P}$ ,  $P_{ij}$  meaning the probability that after rate  $i$  rate  $j$  follows. Another vector,  $\underline{p}$  should contain the probabilities of starting the bursts with given bitrates. If the descriptors of the duration of rate  $i$  burst are  $(\underline{b}_{(i)}, \mathbf{B}_{(i)})$ , then the resultant joint PH distribution, that describes the burst with different transmission rates will have the generator matrix as

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{B}_{(1)} & P_{12} \cdot \underline{E}_{(1)}^0 \cdot \underline{b}_{(2)} & \dots & P_{1R} \cdot \underline{E}_{(1)}^0 \cdot \underline{b}_{(R)} \\ P_{21} \cdot \underline{E}_{(2)}^0 \cdot \underline{b}_{(1)} & \mathbf{B}_{(2)} & \dots & P_{2R} \cdot \underline{E}_{(2)}^0 \cdot \underline{b}_{(R)} \\ \vdots & \vdots & \ddots & \vdots \\ P_{R1} \cdot \underline{E}_{(R)}^0 \cdot \underline{b}_{(1)} & \dots & \dots & \mathbf{B}_{(R)} \end{bmatrix} \quad (2.15)$$

and the initial probability vector is

$$\underline{a}^* = [p_1 \cdot \underline{b}_{(1)} \dots p_R \cdot \underline{b}_{(R)}]. \quad (2.16)$$

## 2.4 Base station model

This Section presents the modelling assumptions applied to describe the behavior of the radio cell or radio base station. In our representation this is the basic element of the network, that provide radio capacity to users and as such, this is the target of performance evaluation presented later. The model presented here does not require the assumption of a particular radio access technology, therefore a base station here is generally modeled as a channel pool of  $C_0$  units of capacity. This capacity is expressed in the same units as the transmission rates of the connections, typically bits per second. However, when using our method during the investigation of a particular system, the capacity may be expressed in other units – although these are also closely related to the rate expressed in bits per second –, for example in TDMA systems the number of time slots in a frame (and users' used capacity is also in this unit), or bandwidth and users' equivalent bandwidth in CDMA systems. As a special case we investigate the scenario when  $C_0 = \infty$  as well, clearly indicating the use and significance of this case.

This kind of session level model is applicable to calculate the rejection probability of newly initiated session or handover attempts. It is clear, that different connection types tolerate rejection differently, e.g. a streaming video session is completely broken when blocked during handover, but the user of a web-browsing session will probably not even notice blocking. Moreover, generally it is less desired to drop a handover attempt than to reject a newly initiated one of the same connection type. Therefore we suppose that several capacity sharing methods are applied at the base station. Namely *Complete Sharing (CS)* policy means that all sessions may enter the system, whenever the unused capacity of the base station is sufficient for handling the session. In the other case channel reservation is applied at the radio interface, implemented by means of a *Partial Sharing (PS)* policy. Practically this means that not all connection types may utilize the whole capacity  $C_0$ , namely there is a maximum available capacity  $C_{k,H}$  and  $C_{k,N}$  for type  $k$  handover and new connections, meaning that a type  $k$  connection can be admitted if the total amount of capacity occupied by type  $k$  sessions after admission is less then or equal to  $C_{k,H}$  or  $C_{k,N}$  and the total occupied capacity does not exceed  $C_0$ . A less general form of Partial Sharing policy is often used in the literature, namely the guard channel concept. This restrains connections that were initiated in the cell from using the total capacity, in favour of handover sessions. Clearly, assuming infinite radio capacity makes the principle of capacity allocation pointless, since there is always enough capacity to serve an arriving customer.

Based on the applied channel sharing policy, the base station may operate according to several admission control mechanisms. Let  $C_{oc}$  denote the amount of instantaneously occupied capacity of the base station at the moment of a session arrival and let  $\underline{c}^{(k)}$  denote the vector containing the possible transmission rates of a type  $k$  connection.

*Burst level CAC* controls arrivals by using instantaneous rate information of sessions. Namely a call is accepted at burst level if  $C_{oc}$  plus its instantaneous rate is less than the allowed capacity. Using symbols, in case of partial sharing, the acceptance of a type  $k$  handover session is assured if

$$C_{oc} + c_j^{(k)} \leq C_0 \quad \text{and} \quad \sum_i n_i^{(H,k)} \cdot c_i^{(k)} + c_j^{(k)} \leq C_{k,H} \quad (2.17)$$

when the customer arrives with rate  $j$ , here  $n_i^{(H,k)}$  is the number of type  $k$  handover customers transmitting with rate  $i$  at the moment. Burst blocking occurs if (2.17) does not hold. In some cases burst blocking would not cause the breakdown of the session, but while the connection is maintained it suffers some degradation of packet level QoS measures (for instance increased queueing delays or dropped packets). From the users point of view it appears as some "distur-

bance" in the transmission, for example missing or incomprehensible periods of speech, "broken up" or stilled pictures of video, or suddenly decreased resolution of images. To model maintained, but temporarily degraded connections due to burst blocking, we introduce the *rate reduction* policy of burst level admission control. If rate reduction is applied, upon arrival of a type  $k$  handover flow and if (2.17) does not hold the connection is forced to reduce its rate to a level that "fits into" the channel. To conclude, within burst level CAC two policies can be considered, depending on the type of arriving connection:

- policy 1, *immediate blocking*: the connection is immediately blocked. In case of handover connection or blocking sensitive connection type this is not tolerable.
- policy 2, *rate reduction*: the connection is forced to reduce its transmission rate. Assuming again an arriving handover session and partial sharing, if  $c_j^{(k)}$  is the highest transmission rate so that

$$C_{oc} + c_j^{(k)} \leq C_0 \quad \text{and} \quad \sum_i n_i^{(H,k)} \cdot c_i^{(k)} + c_j^{(k)} \leq C_{k,H},$$

the connection is forced to begin its transmission with rate  $c_j^{(k)}$ . The connection is only blocked when

$$C_{oc} + c_{min}^{(k)} > C_0 \quad \text{or} \quad \sum_i n_i^{(H,k)} \cdot c_i^{(k)} + c_{min}^{(k)} > C_{k,H},$$

where  $c_{min}^{(k)}$  is the lowest possible transmission rate. Again, applying this policy may result in the degradation of packet level QoS.

It is worth noting that these two approaches can be found in the literature, sometimes named full and partial blocking (e.g. [11]).

It is clear that when sources are characterized by possible zero transmission rate and the probability of arriving with zero rate is not zero, burst level CAC does not bound the maximum number of admissible sessions. Applying the rate reduction policy may accept even more connections.

Because the connections change their transmission rate during the session, the amount of occupied capacity at the base station may change without the arrival or termination of a connection. This means that burst blocking of an already admitted connection may occur during transmission. This happens when a flow tries to switch to a transmission rate with which the total occupied capacity would exceed the maximum available.

# Chapter 3

## Queueing model of the radio cell

In this Chapter the queueing model of the formerly described cellular system is derived and investigated. In order to utilize well known queueing theory methods and to keep the model general and comprehensive I formerly supposed that the customer describing times have phase type distributions and I have shown the viability of this assumption. Based on this, an approximate formula is derived, that enables the fast and efficient computation of session level performance parameters with reasonable error.

### 3.1 Service process

This Section is devoted to present the notion and derivation of the parameters of the service process applicable in the queueing model representing the cellular environment described in the previous Chapter. In order to formulate this queueing model, a service process is needed that has the following properties:

- the service time distribution is equivalent with the channel occupancy time distribution, it satisfies (2.6),
- the service process describes the instantaneous transmission rate of the customer.

In order to define the service process with the above properties, first we compose the channel holding time distribution for type  $k$  new connections, using the distributions of the residual dwell time and the session duration. The same method should be used when determining the channel occupancy time distribution for type  $k$  handover sessions, using the dwell time and residual session length distributions.

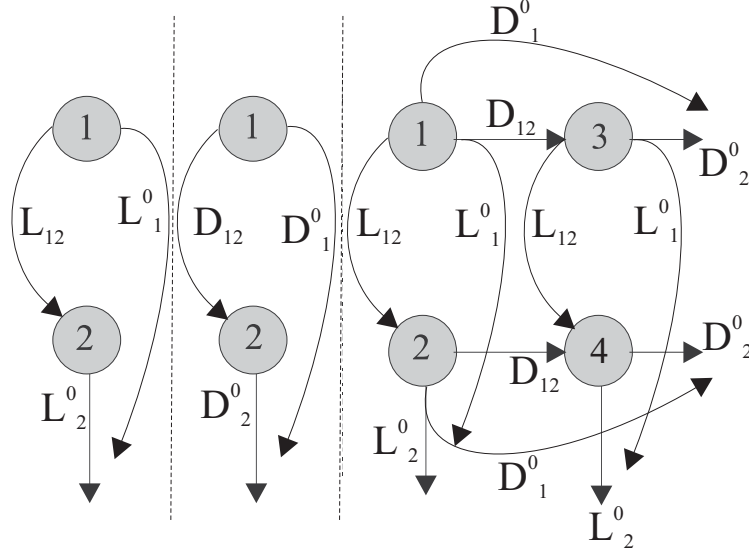


Figure 3.1: Construction of the channel holding time distribution

Let the phase number of the residual dwell time of a type  $k$  customer be denoted by  $N_{D,R}^k$ , that of the connection duration is denoted by  $N_L^k$ . Then the channel holding time also has a phase type distribution with  $N_{D,R}^k \cdot N_L^k$  phases and descriptors  $(\mathbf{T}^{(N,k)}, \underline{t}^{(N,k)})$ . This distribution is composed as follows. We form  $N_{D,R}^k$  groups representing the phases of the residual dwell time, each containing  $N_L^k$  phases. Among the phases of a group the rates are equal to the rates of the phase type distributed session length. Among those phases of different groups that represent the same phase of the session duration the rates are equal to the rates of the residual dwell time distribution. This means that the rate between phase  $i$  of group  $n$  and phase  $j$  of group  $m$  is:

- $L_{ij}^{(k)}$  if  $n = m$ ,  $i \neq j$ ,
- $D_{nm}^{(R,k)}$  if  $i = j$ ,  $n \neq m$
- 0 if  $i \neq j$ ,  $n \neq m$ ,

for  $i, j = 1, \dots, N_L^k$ ,  $n, m = 1, \dots, N_{D,R}^k$ .

If the phases of the channel holding time are enumerated appropriately, the initial probability vector and generator matrix of its distribution are given as:

$$\underline{t}^{(N,k)} = \underline{d}^{(R,k)} \otimes \underline{l}^{(R,k)} \quad \mathbf{T}^{(N,k)} = \mathbf{D}^{(R,k)} \oplus \mathbf{L}^{(k)} \quad (3.1)$$

where  $\otimes$  and  $\oplus$  denotes the Kronecker product and Kronecker sum of two matrices. An example of the composition of the channel holding time is presented on Figure 3.1. Here both the dwell

time and session length has a 2 phase Coxian distribution, the right hand side of the Figure contains the 4 phase channel holding time distribution with its rates. First we have to see that this distribution satisfies (2.6).

**Theorem 3.1.1.** *Let  $x, y$  be independent phase type distributed random variables with distributions characterized by  $(\underline{d}, \mathbf{D})$ ,  $(\underline{l}, \mathbf{L})$ . If  $z$  is a PH distributed random variable with descriptors  $(\underline{t}, \mathbf{T})$  and (3.1) holds, then  $z = \min(x, y)$ .*

**Proof.** *Let the distribution function of  $x$  and  $y$  be denoted by  $X(t)$  and  $Y(t)$ . The distribution function of  $\min(x, y)$  is then  $M(t) = 1 - (1 - X(t))(1 - Y(t))$ . Substituting the cdf of PH distributions from (2.7), we get*

$$M(t) = 1 - \underline{d}e^{\mathbf{D}t}\underline{h}_{N_D} \cdot \underline{l}e^{\mathbf{L}t}\underline{h}_{N_L} \quad (3.2)$$

where  $N_D$  and  $N_L$  denote the number of phases of the distributions of  $x$  and  $y$ ,  $\underline{h}_{N_D}$  and  $\underline{h}_{N_L}$  are column vectors of length  $N_D$  and  $N_L$  filled with 1s. Introducing  $\underline{a}(t) = \underline{d}e^{\mathbf{D}t}$  and  $\underline{b}(t) = \underline{l}e^{\mathbf{L}t}$ , (3.2) has the form:

$$M(t) = 1 - \left( \sum_{i=1}^{N_D} a_i(t) \right) \left( \sum_{i=1}^{N_L} b_i(t) \right).$$

The distribution function of  $z$  is:

$$Z(t) = 1 - \underline{t}e^{\mathbf{T}t}\underline{h}_T = 1 - (\underline{d} \otimes \underline{l})e^{(\mathbf{D} \oplus \mathbf{L})t}\underline{h}_{N_D \cdot N_L}.$$

Using the properties of the Kronecker operations<sup>1</sup> the expression is transformed to

$$Z(t) = 1 - ((\underline{d}e^{\mathbf{D}t}) \otimes (\underline{l}e^{\mathbf{L}t})) \underline{h}_{N_D \cdot N_L}.$$

Substituting  $\underline{a}(t)$ ,  $\underline{b}(t)$  and using the definition of the Kronecker product, we get

$$Z(t) = 1 - \sum_{i=1}^{N_D} a_i(t) \cdot \left( \sum_{j=1}^{N_L} b_j(t) \right) = 1 - \left( \sum_{i=1}^{N_D} a_i(t) \right) \cdot \left( \sum_{j=1}^{N_L} b_j(t) \right) = M(t). \quad \square$$

To include the instantaneous transmission rate of a customer into the service process, a similar method is necessary, using the PH distributed channel occupancy time and the rate describing

---

1

- $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ ,
- $e^{\mathbf{A} \oplus \mathbf{B}} = e^{\mathbf{A}} \otimes e^{\mathbf{B}}$ .

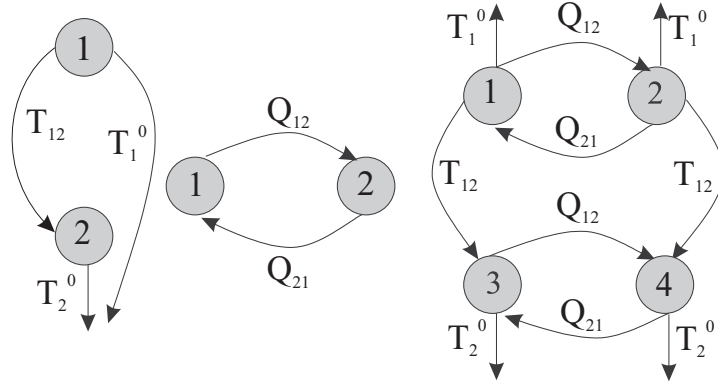


Figure 3.2: Construction of the service time distribution

underlying Markov chain. In this case the states of the traffic describing Markov chain are organized in as many groups as many phases the channel holding time distribution has. Among the states of the same group the rates are equal to the rates within the state structure of the general Markovian model. Between two states of two different groups that are in the same position within their group the appropriate rate of the channel holding time distribution characterizes the phase transition. Following its construction, we arrive to the service process, that is described by a PH distribution, with the following parameters

$$\underline{s}^{(N,k)} = \underline{t}^{(N,k)} \otimes \underline{q}^{(k)} \quad \mathbf{S}^{(N,k)} = \mathbf{T}^{(N,k)} \oplus \mathbf{Q}^{(k)}. \quad (3.3)$$

The duration of this service process is equivalent with the channel holding time duration, while the actual state is assigned to the actual state of the transmission rate generator Markov chain.

Figure 3.2 shows an example of constructing the service process. Here the channel holding time is presented with 2 phases and the rate describing Markov chain has 2 states. The right hand side of the Figure depicts the service time distribution with its appropriate rates. As we can see, this construction of the service process is analogous with the construction of the active period model presented as part of the general Markovian source model in Section 2.3, equation (2.11), thus the following theorem applies for the active period model as well. Now we have to show that this service time has the same distribution as the channel holding time.

**Theorem 3.1.2.** *Let  $x$  denote a PH distributed random variable with descriptors  $(\underline{t}, \mathbf{T})$  and probability density function  $f(t)$ . Let  $\mathbf{Q}$  and  $\underline{q}$  denote the infinitesimal generator matrix and the initial probability vector of a finite state CTMC. If  $y$  is a PH distributed random variable with descriptors  $(\underline{s}, \mathbf{S})$  and (3.3) holds, then its pdf is  $f(t)$ .*

**Proof.** Following the construction of the distribution of  $y$ , it is obvious that the column vector  $\underline{S}^0$  containing the finishing rates from all the phases of this distribution is expressed as

$$\underline{S}^0 = \underline{T}^0 \otimes \underline{h}_{N_Q},$$

where  $N_Q$  denotes the number of states of the Markov chain and  $\underline{T}^0$  is the column vector containing the finishing rates from each phase of the distribution of  $x$ .

Substituting (3.3) into the density function of  $y$  according to (2.7) we get:

$$g(t) = (\underline{t} \otimes \underline{q}) e^{(\mathbf{T} \oplus \mathbf{Q})t} \left( \underline{T}^0 \otimes \underline{h}_{N_Q} \right).$$

Using the properties of the Kronecker product the expression changes:

$$g(t) = (\underline{t} \otimes \underline{q})(e^{\mathbf{T}t} \otimes e^{\mathbf{Q}t})(\underline{T}^0 \otimes \underline{h}_{N_Q}) = (\underline{t}e^{\mathbf{T}t}) \otimes (\underline{q}e^{\mathbf{Q}t})(\underline{T}^0 \otimes \underline{h}_{N_Q}) = (\underline{t}e^{\mathbf{T}t}\underline{T}^0) \otimes (\underline{q}e^{\mathbf{Q}t}\underline{h}_{N_Q}).$$

The second argument of the last Kronecker product is always 1, since  $\underline{q}e^{\mathbf{Q}t}$  is a probability vector. Thus the equation reduces to:

$$g(t) = \underline{t}e^{\mathbf{T}t}\underline{T}^0 \otimes 1 = \underline{t}e^{\mathbf{T}t}\underline{T}^0 = f(t). \quad \square$$

This composition of the service process allows us to determine the instantaneous transmission rate of a connection as well as the phase of the dwell time and the session length distributions. Namely if a new session is in the  $(m-1) \cdot N_L^{(k)} \cdot N_D^{(R,k)} + (j-1) \cdot N_L^{(k)} + i$ th phase of the service time, its transmission rate is  $c_m^{(k)}$  and the actual phase of its session length and residual dwell time is  $i$  and  $j$  respectively. The service time of a type  $k$  new session has  $N_L^{(k)} \cdot N_D^{(R,k)} \cdot N_Q^{(k)}$  phases which causes that the Markov chain model of the system has multiple dimensions and possibly huge state space.

It is also clear from the composition of the service process that the transmission rate of a connection in those phases of the service time that correspond to a particular state of the traffic describing Markov chain is identical. Namely if  $\underline{r}^{(N,k)}$  denotes the vector containing the transmission rates  $r_i^{(N,k)}$  of a type  $k$  new connection if the session is receiving the  $i$ th phase of the service time ( $\underline{r}^{(H,k)}$  is the same for handover connections), then

$$\begin{aligned} r_1^{(N,k)} &= r_2^{(N,k)} = \dots = r_{N_L^{(k)} \cdot N_D^{(R,k)}}^{(N,k)} = c_1^{(k)} \\ r_{N_L^{(k)} \cdot N_D^{(R,k)} + 1}^{(N,k)} &= r_{N_L^{(k)} \cdot N_D^{(R,k)} + 2}^{(N,k)} = \dots = r_{2N_L^{(k)} \cdot N_D^{(R,k)}}^{(N,k)} = c_2^{(k)} \\ &\vdots \\ r_{(N_Q^{(k)} - 1) \cdot N_L^{(k)} \cdot N_D^{(R,k)} + 1}^{(N,k)} &= \dots = r_{N_Q^{(k)} \cdot N_L^{(k)} \cdot N_D^{(R,k)}}^{(N,k)} = c_{N_Q^{(k)}}^{(k)}. \end{aligned} \quad (3.4)$$

The above construction of the service process should be carried out for each customer type  $k, k \in \{1 \dots K\}$ . Moreover, handover and new connections of the same type are different in terms of their describing times and the initial probability vector of the traffic characterising Markov chain, thus each customer type requires the use of two service time distributions. This results in the composition of  $2K$  service time distributions. In this proposed model we do not allow a customer to change its service type during a session, thus the service times could be handled independently.

## 3.2 The driving process

In this section the Markov process that describes the cellular model is presented. Given the incoming process of new and handover sessions is supposed to be Poissonian, the service time is PH and customers change their amount of occupied capacity during service formally the system is described by a multiclass M/PH/ $C_0$  queue with phase dependent capacity requirements.

The state of the resulting Markov process is the vector

$$\underline{n} = [\underline{n}^{(N,1)}, \dots, \underline{n}^{(N,K)}, \underline{n}^{(H,1)}, \dots, \underline{n}^{(H,K)}], \quad (3.5)$$

where the  $i$ th element of vector  $\underline{n}^{(N,k)}$ ,  $n_i^{(N,k)}$  denotes the number of type  $k$  sessions that arrived to the cell as new connection and receiving the  $i$ th phase of the type  $k$  new connection service time,  $\underline{n}^{(H,k)}$  contains the same quantities regarding handover connections.

Considering the capacity reservation described in Section 2.4 and the transmission rates introduced in (3.4) we may determine the boundaries of the state space. Namely the valid states of the system are those  $\underline{n}$ -s, where

$$\begin{aligned} \underline{n}^{(N,k)} \cdot \underline{r}^{(N,k)} &\leq C_{k,N}, & \underline{n}^{(H,k)} \cdot \underline{r}^{(H,k)} &\leq C_{k,H}, & k &= 1, \dots, K \\ \sum_{k=1}^K \underline{n}^{(N,k)} \cdot \underline{r}^{(N,k)} + \sum_{k=1}^K \underline{n}^{(H,k)} \cdot \underline{r}^{(H,k)} &\leq C_0 \end{aligned} \quad (3.6)$$

holds. This simply means that the amount of occupied capacity of each type new and handover connections may not exceed the maximum available for that type and the total amount of used capacity cannot be larger than the maximum available at the base station.

The state space of the system is finite if the connections always transmit with some rate. This means that none of the traffic describing Markov chains contains a state that is associated with zero transmission rate. In this case the system is irreducible and finite, hence it is stable and its

steady state distribution exists. If zero transmission is allowed (e.g. the well known and widely used on-off sources) the state space becomes infinite. However, regardless a session's transmission state it stays in the system until the channel occupancy time and since no restrictions were given on the number of customers in the cell from this point of view the cell acts as an infinite server queue, thus the criteria for stability is to have incoming rates finite, that is completed in our case with the requisite of having finite rates of the traffic describing Markov chain.

To unambiguously describe the Markov chain representing the proposed cellular system, the possible state transitions and the rates of these transitions are needed. State transitions may occur because of the following events: a handover or new connection arrives to the system, a session leaves the system either by handover or by connection termination or a customer changes the phase of its service time, this latter obviously includes the change of transmission rate of a customer. To simplify notations in the following discussion we do not use the whole state vector  $\underline{n}$ , rather its subvector  $\underline{n}^{(N,k)}$  or  $\underline{n}^{(H,k)}$  that changes due to a state transition. The state transition rates are also dependent on the applied admission control policy of the base station described in Section 2.4.

Recalling Section 2.1 the incoming rate of handover and new connections is denoted by  $\lambda_H$  and  $\lambda_N$ , the probability that a particular arriving customer is of type  $k$  is denoted by  $\alpha_k$ . If for type  $k$  admission control policy 1 is applied the state transition rates are the following:

- state transition due to a new connection arrival: this event results in a state transition from state  $\underline{n}^{(N,k)}$  into state  $\underline{n}^{(N,k)} + \underline{e}_i$  at rate  $\lambda_N \cdot \alpha_k \cdot s_i^{(N,k)}$ , where  $\underline{e}_i$  is a vector of the same size as the phase number of the service time of type  $k$  connections, filled with 0's and one 1 at its  $i$ th position;
- state transition due to handover arrival: this results in a state transition from state  $\underline{n}^{(H,k)}$  into state  $\underline{n}^{(H,k)} + \underline{e}_i$  at rate  $\lambda_H \cdot \alpha_k \cdot s_i^{(H,k)}$ ;
- state transition due to session termination: this event results in a transition from state  $\underline{n}^{(N,k)}$  into state  $\underline{n}^{(N,k)} - \underline{e}_i$  at rate  $n_i^{(N,k)} \cdot S_i^{(N,k,0)}$ , where the vector  $\underline{S}^{(N,k,0)}$  contain the finishing rates from all the phases of the type  $k$  new connection service time;
- state transition due to the phase change of the service time distribution: this results in a transition from state  $\underline{n}^{(N,k)}$  into state  $\underline{n}^{(N,k)} - \underline{e}_i + \underline{e}_j$  at rate  $n_i^{(N,k)} \cdot S_{ij}^{(N,k)}$ .

Obviously the above transitions may occur when the total occupied capacity after transition does not exceed the amount that is allowed for a particular connection type, namely in case of an

arrival with transmission rate  $c_i^{(k)}$

$$C_{oc} + c_i^{(k)} \leq C_{k,N} \quad \text{or} \quad C_{oc} + c_i^{(k)} \leq C_{k,H},$$

or in case of phase change that result in a change of transmission rate from  $c_i^{(k)}$  to  $c_j^{(k)}$

$$C_{oc} - c_i^{(k)} + c_j^{(k)} \leq C_{k,N} \quad \text{or} \quad C_{oc} - c_i^{(k)} + c_j^{(k)} \leq C_{k,H},$$

where again  $C_{oc}$  denotes the amount of total occupied capacity prior to state transition.

To classify the transition rates when admission control policy 2 (burst CAC, rate reduction) is applied, we suppose without the loss of generality that the possible transmission rates (so the states of the traffic describing Markov chain as well) are enumerated in increasing order from the lowest rate to the highest. It is clear from the construction of the service time distribution that for those phases  $i$  and  $j$  of the service time distribution that correspond to different states of the rate describing Markov chain, but to the same phase of the channel holding time distribution

$$|i - j| = l \cdot N_D^{(R,k)} \cdot N_L^{(k)} \quad l \in [1, \dots, N_Q^{(k)} - 1]$$

holds. In the case when policy 2 is applied and a connection is forced to reduce its transmission rate due to the lack of capacity, the initial phase of its service time is altered to a phase that corresponds to a lower transmission rate but to the same phase of the channel occupancy time. Supposing that phase  $i$  of the service time distribution corresponds to the highest admissible transmission rate the state transition rates due to arrival has the form:

- a new or handover connection results in a state transition from state  $\underline{n}^{(N,k)}$  or  $\underline{n}^{(H,k)}$  into state  $\underline{n}^{(N,k)} + \underline{e}_i$  or  $\underline{n}^{(H,k)} + \underline{e}_i$  at rate

$$\lambda_N \cdot \alpha_k \cdot \sum_{l=0}^{N_Q^{(k)} - \left\lfloor \frac{j}{N_Q^{(k)}} \right\rfloor - 1} s_{i+l \cdot N_D^{(R,k)} \cdot N_L^{(k)}}^{(N,k)} \quad \text{and} \quad \lambda_H \cdot \alpha_k \cdot \sum_{l=0}^{N_Q^{(k)} - \left\lfloor \frac{j}{N_Q^{(k)}} \right\rfloor - 1} s_{i+l \cdot N_D^{(R,k)} \cdot N_L^{(k)}}^{(H,k)}.$$

By observing the difference between the incoming rates of the two investigated admission control policies it is clear that policy 2 allows more connections to be admitted in case of an overloaded base station. After all, although this policy generally reduces blocking probabilities, reducing the transmission rate of a connection may not be realized for some type of connections due to the resulting degradation of packet level QoS measures.

After having determined all the possible transition rates among system states, calculating the steady state distribution of the system is the straight way to achieve session blocking probabilities and channel utilization values. We assume Poissonian arrival of sessions and these type of arrivals have the well-known property of seeing the time average (PASTA property – Poisson Arrivals See Time Average), e.g. stationary distribution of the system ([98]). Thus determining blocking probabilities means summing up the probabilities of those states in which session rejection may occur due to the lack of base station capacity.

To determine the steady state distribution of the system we must enumerate the states and compose the infinitesimal generator matrix of the system using the above determined possible transmissions and their rates. Given the infinitesimal generator  $Q^*$  (note that it is not the same as the infinitesimal generator of the Markov chain describing the variability of the transmission rate of a customer, formerly denoted by  $Q^{(k)}$ ), the steady state probability vector  $\underline{p}^*$  is calculated by solving the well known set of global balance equations, i.e.

$$\underline{p}^* Q^* = 0 \quad \underline{p}^* \underline{h} = 1.$$

Unfortunately in practical scenarios solving this system of equations is not possible because of the huge number of states. In very simple scenarios the state space may contain only a few hundred thousand states. In this case the system of equations may be solved by some numerical method (e.g. Gauss-Seidel). But in most cases the number of states easily exceeds tens of millions, than solving the system of global balance equations is impossible. Moreover, if a session may arrive with zero transmission rate, the state-space becomes infinite, thus this approach is not applicable as well.

Fortunately, to determine blocking behavior and system utilization, we do not need the steady state distribution itself. Rather it is enough if the probability of having  $m$  units of system capacity occupied is known,  $m = 1, \dots, C_0$ . These parameters are referred as channel occupancy probabilities in the remainder of this Chapter.

### 3.3 Local balance and product form

This Section presents the local balance and product form steady state distribution that holds in case of infinite capacity assumption; this will be used later for deriving the approximate analysis. To develop a fast and efficient method of calculating channel occupancy probabilities of the system we have to take a deeper insight of the balance equations of the system describing

Markov chain. To do this, we consider the theoretical case when the base station has infinite capacity, i.e.  $C_0 = \infty$ . Moreover, this case has not only theoretical, but practical significance in teletraffic modeling of systems, namely when system performance is determined in terms of *overload probability*, that is the probability of total occupied capacity exceeding a given value. This approach is used in the literature when dealing with CDMA networks, where capacity does not have a hard bound, but overload can be well defined.

It is clear, that when the base station is supposed to have infinite capacity, the Markov chain has a product form solution. This is because the infinite capacity case is equivalent with a queueing network, if each phase  $i$  of the service time distribution is modelled by a single infinite server exponential queue with rate  $-S_{ii}$  and with routing probability  $\frac{S_{ij}}{-S_{ii}}$  between queues  $i$  and  $j$ . We know from the famous BCMP theorem [99], that this network has a product form equilibrium distribution, containing the product of the steady state distribution of individual queues. In the context of queueing networks, this is the usual use of the term product form solution. However, as the system under consideration in this thesis is a single queue we use the term in more general sense. That is: product form distribution is a multi-dimensional distribution that is the product of its marginal distributions (e.g. [100]). Naturally, this latter more general definition contains the stationary distributions of product form queueing networks, as the distribution of the queue length of individual queues are the marginal distributions of the joint queue lengths distribution.

The system's analogy with the BCMP network containing IS queues also means that this system is quasi-reversible and local balance equations hold (e.g. [101]). We now that in BCMP networks the local balance equations equate the input and output rates of a network station. Obviously, in our problem the non-blocking part of the state space (i.e. those states, where any state transition can take place, the idle capacity of the base station is enough to accommodate a new connection with any instantaneous transmission rate) is equivalent with the state space of the infinite capacity case, in terms of the possible state transitions and their rates. Thus, in the non-blocking part the local balance equations will also hold. We will use the local balance equations to determine the closed form equilibrium distribution (product form) and later use it for approximate analysis as well. The BCMP analogy means that the following transitions hold balance in our system:

- transitions that result in an increment of the number of customers receiving a particular phase of the service time, caused by an arrival of a new or handover session, or by the phase change of an active customer

- transitions that result in the diminution of the number of customers receiving the same phase of the service time, caused by one customer leaving the system (either by handover or by connection termination), or by phase change of a customer.

Since no transition is allowed among different customer types, the following equations are valid for any type  $k$  sessions initiated within the examined cell or arrived after handover. The following derivation considers type  $k$  new connections, but for the sake of better readability the corresponding part of the state vector, i.e. vector  $\underline{n}^{(N,k)}$  is simply denoted by  $\underline{n}^*$ . Moreover, for better readability we omit the superscript  $(N, k)$  of the corresponding service time matrix and initial vector as well. Putting all these into the language of mathematical symbols, the local balance equations have the form:

$$\lambda_N \alpha_k s_i p(\underline{n}^*) + \sum_{j=1, j \neq i}^P (n_j^* + 1) S_{ji} \cdot p(\underline{n}^* + \underline{e}_j) = (n_i^* + 1) \cdot p(\underline{n}^* + \underline{e}_i) \left( S_i^0 + \sum_{j=1, j \neq i}^P S_{ij} \right), \quad (3.7)$$

where  $\underline{e}_i$  denotes a vector of length  $P$  filled with zeros and a 1 in its  $i$ th position and for the sake of simplicity the number of phases of the service time of type  $k$  new sessions  $N_{D,R}^{(k)} \cdot N_L^{(k)} \cdot N_Q^{(k)}$  is denoted by  $P$ . Although these local balance equations are determined using the system's identity with a BCMP network, an alternative proof of the validity of (3.7) equations can be found in Appendix C.

Rearranging (3.7) and writing all the local balance equations into a single vectorial equation, we get

$$-\lambda_N \alpha_k \underline{s} \cdot p(\underline{n}^*) = [(n_1^* + 1)p(\underline{n}^* + \underline{e}_1), \dots, (n_i^* + 1)p(\underline{n}^* + \underline{e}_i), \dots, (n_P^* + 1)p(\underline{n}^* + \underline{e}_P)] \mathbf{S}. \quad (3.8)$$

Introducing the vector

$$\underline{F} = \left( \frac{(n_1^* + 1)p(\underline{n}^* + \underline{e}_1)}{p(\underline{n}^*)}, \dots, \frac{(n_P^* + 1)p(\underline{n}^* + \underline{e}_P)}{p(\underline{n}^*)} \right) \quad (3.9)$$

from (3.8) we have

$$\underline{F} = -\lambda_N \alpha_k \underline{s} \mathbf{S}^{-1}. \quad (3.10)$$

We can see that the relation of the steady state probabilities of neighboring states is contained in this vector, namely

$$F_i \cdot p(\underline{n}^*) = (n_i^* + 1) \cdot p(\underline{n}^* + \underline{e}_i). \quad (3.11)$$

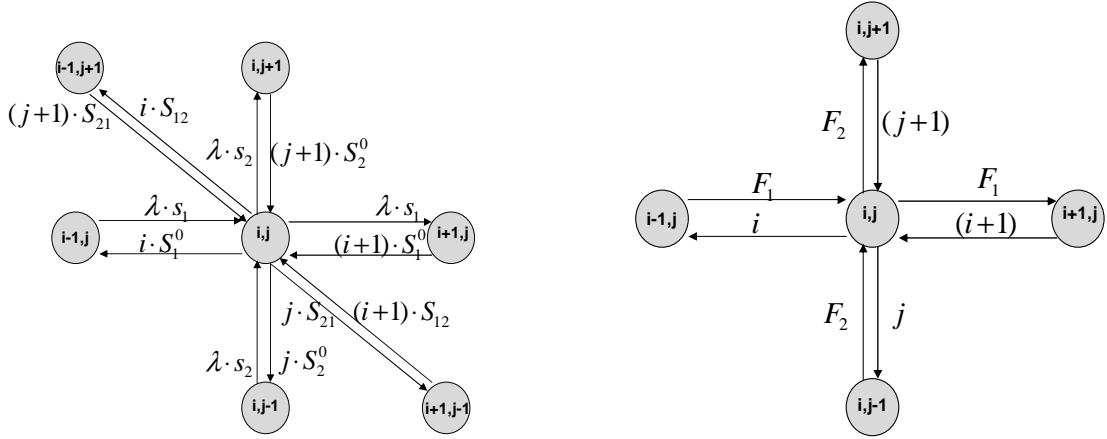


Figure 3.3: Two dimensional state space example

This indicates that the equilibrium probability of state  $\underline{n}^*$  in the infinite capacity case can be calculated recursively from state  $\underline{0}$ , by means of iteratively substituting into (3.11). This results in the steady state probability of  $\underline{n}^*$  as

$$p(\underline{n}^*) = p(\underline{0}) \cdot \prod_{i=1}^P F_i^{n_i^*} \cdot \frac{1}{n_i^*!}. \quad (3.12)$$

An alternative derivation of (3.12), based on the BCMP analogy can be found in Appendix C. We observe that (3.12) is identical with the steady state distribution of a multiclass M/G/m/m or M/G/ $\infty$  system (e.g. [102]), with  $P$  classes and with offer loads  $F_i$  for class  $i$  (no change of required capacity during a connection). Figure 3.3 shows an example of a two dimensional system, namely the state space around state  $(i, j)$ . This state space would describe a system with single session type that has two possible transmission rates and exponential channel occupancy time. Only those transitions and states are plotted that affect state  $(i, j)$ . The right hand side of the figure shows the equivalent state space: here the elements of  $\underline{F}$  replace the arrival rates, while the leaving rates are normalised to 1. We see that with this description based on the local balance equations the "crossing" transitions disappeared and their influence is included in  $\underline{F}$ . As the elements of  $\underline{F}$  characterise the ratio of the stationary probabilities of two neighboring states (3.11), I use the notion of *multiplier factor* for the elements of  $\underline{F}$  and *multiplier vector* for  $\underline{F}$ . Now, returning to the original, complete notions of this thesis, the complete and correct form of the multiplier vector for a type  $k$  new connection is:

$$\underline{F}^{(N,k)} = -\lambda_N \alpha_k \underline{S}^{(N,k)} \left( \underline{S}^{(N,k)} \right)^{-1}. \quad (3.13)$$

The complete expression of the product form distribution defined in (3.12) is [99]

$$p(\underline{n}) = \frac{1}{G} \prod_{k=1}^K \left( \prod_{i=1}^{P(N,k)} \left( F_i^{(N,k)} \right)^{n_i^{(N,k)}} \frac{1}{n_i^{(N,k)}!} \prod_{i=1}^{P(H,k)} \left( F_i^{(H,k)} \right)^{n_i^{(H,k)}} \frac{1}{n_i^{(H,k)}!} \right) \quad (3.14)$$

where  $P^{(N,k)}$  denotes the number of phases of type  $k$  new customers' service time and  $G$  is the normalization constant so that the sum of probabilities of all states equals to one. Namely

$$\sum_{\text{all } \underline{n}} p(\underline{n}) = 1 \quad \Rightarrow \quad G = \sum_{\text{all } \underline{n}} \prod_{k=1}^K \left( \prod_{i=1}^{P(N,k)} \left( F_i^{(N,k)} \right)^{n_i^{(N,k)}} \frac{1}{n_i^{(N,k)}!} \prod_{i=1}^{P(H,k)} \left( F_i^{(H,k)} \right)^{n_i^{(H,k)}} \frac{1}{n_i^{(H,k)}!} \right). \quad (3.15)$$

Recognizing that the infinite capacity assumption allows the sum to run from 0 to  $\infty$  for all  $n_i$ -s, the sum is the Taylor series of the exponential function, namely

$$G = \exp \left( \sum_{k=1}^K \left( \sum_{i=1}^{P(N,k)} F_i^{(N,k)} + \sum_{i=1}^{P(H,k)} F_i^{(H,k)} \right) \right). \quad (3.16)$$

In practical cases the number of phases of the service time distribution may be quite large, especially if exact PH fitting with numerous phases (e.g. 10-12) is used for the session length and dwell time distributions and sophisticated user traffic model is used. In this case the matrix inversion of (3.13) is the most resource demanding task.

### 3.4 Approximate analysis of the finite capacity case

This Section presents the approximate formula that allows the determination of the distribution of channel occupation, which is the base for determining performance measures. Now let us consider the state space of the system proposed, in particular the subspace that contains states where the highest capacity demands cannot be admitted, because customers using lower rates occupy the capacity. To visualise this situation, I consider a simple system with 20 units of capacity and users might transmit using 1, 3 or 10 units. If we consider the channel occupancy distribution to be exponential, this system has three dimensional state space, each dimension representing the number of customers instantaneously using a given amount of unit from the three possibilities. The state space of this example is shown in Figure 3.4. Here the abscissa is the number of users transmitting with 1, the ordinate is the number of customers using 3 capacity and the overlapping rectangular means one customer using 10 units (one level "up" along the  $z$

coordinate), the overlapping triangle means two customers using 10 units (two levels "up" along the  $z$  coordinate).

The subspace under consideration consists of the states denoted by single circles and not side neighbors of the plotted rectangles: these are the states where there is zero customers using 10 units and the switching to 10 units or arrival with 10 units instantaneous requirement is not allowed because of capacity limitations (in the states that are side-neighbors of the plotted rectangles, switching to the highest rate is still possible, but arrival with the highest rate is not), in the Figure these are the states over the dashed line. It is obvious that the transitions within this subspace are governed by the matrix of the service time distribution, without those entries that describe the transition rates into or from phases that would mean switching to or from the highest capacity demand. The arrivals are characterised by the initial probability vector of the service time distribution as well, with those entries set to zero that refer to arrival to the phase with highest capacity demand.

So in terms of arrival rates, finishing rates and phase change rates this subspace is governed by the same state transitions as if we were considering a system where the highest transmission rate does not exist. If we consider a system with infinite capacity and a service time distribution that lacks those phases that correspond to the highest instantaneous capacity requirement, naturally all the findings of the previous Section would hold (local balance equations, product form stationary distribution) for this system. This inspires the derivation of the multiplier factors for this system, which is naturally calculated according to (3.10), with the new versions of  $\underline{s}$  and  $\underline{S}$  (not containing entries corresponding to phases with highest transmission rate). This train of thought can be further continued, considering the subspace where the second largest transmission rates also cannot be admitted or switched to, the analogous system where the two largest transmission rates do not exist, derivation of corresponding multiplier factors, etc. In the previous example (Figure 3.4) this means the two states at the right end of the Figure.

Actually, in the subspace where new arrival with and switching to the highest capacity requirement is not allowed (blocking subspace) – note that this general definition contains more states we considered above, namely those states as well, where the number of customers using the highest capacity requirement is not necessarily zero, e.g. the states denoted by those rectangles in Figure 3.4 that are not side neighbors of the triangle –, the system behaves as if the service time distribution changed. This change depends on the instantaneous occupied capacity at that state and the allowed capacity for the given customer type.

Formulating the change of the service time of a type  $k$  new call, when admission control

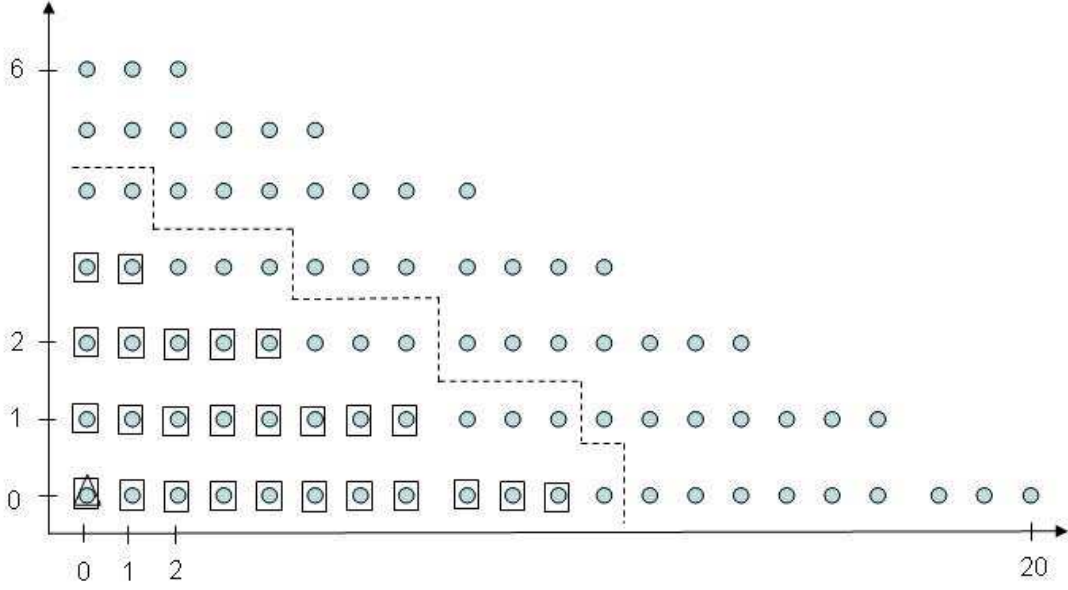


Figure 3.4: Two dimensional state space example

policy 1 (immediate blocking) is applied and  $x$  units of capacity is occupied we get

$$\mathbf{S}_{ij}^{(k)}(x) = 0, \quad s_j^{(k),N}(x) = 0, \quad \forall j : r_j^{(k)} > C_{k,N} - x. \quad (3.17)$$

The diagonal elements of the rate matrix are updated so that  $\sum_j \mathbf{S}_{ij}^{(k)}(x) + S_i^{(k),0} = 0$ . For handover connections the rate matrix and the initial probability vector is changed analogously.

If the second admission control policy is applied, the rate matrix changes the same way as for policy 1. If  $r_i^{(k)}$  denotes the highest transmission rate such that  $r_i^{(k)} \leq C_{k,N} - x$  the initial probability vector changes as:

$$s_i^{(k),N}(x) = \sum_{j: r_j^{(k)} > C_{k,N} - x} s_j^{(k),N} + s_i^{(k),N}, \quad (3.18)$$

where  $s_j^{(k),N}$  means the  $j$ th element of the original initial probability vector.

The multiplier factor defined by (3.13) is calculated from the parameters of the service time distribution. As we have seen, in blocking subspaces the system behaves as if the service time distribution changed, it is straightforward to introduce to load dependent multiplier factor, that is calculated from the changed service time distribution, namely

$$\underline{F}^{(k),N}(x) = -\lambda_N \cdot \alpha_k \cdot \underline{s}^{(k),N}(x) \cdot (\mathbf{S}^{(k)}(x))^{-1}. \quad (3.19)$$

To derive an approximate formula for investigating the system described above, we should look at the work of Kaufman [103] and Roberts [104]. The problem they addressed is the calculation of blocking probabilities on a shared channel (i.e. radio interface, high speed link), without the calculation of the steady state distribution of the Markov chain modeling the shared channel. Their model included several customer types, with different but constant capacity requirements, Poisson arrivals and arbitrary holding time distributions sharing a common channel. This model is the multiclass M/G/m/m queue that has a product form steady state distribution [102], in the form of (3.12). The basic idea introduced was a mapping of the system describing Markov chain, from the multi-dimensional state space into a one-dimensional space, where the total amount of occupied capacity is followed. Then a recursive solution is given, to calculate the channel occupancy probabilities and this recursion uses the multiplier factors ( $F_i$ ) that occur in the product form solution. From these channel occupancy probabilities the blocking probabilities and channel utilisation are easy to calculate.

The original Kaufman-Roberts recursive formula is applicable to systems having product form stationary distributions as (3.12). The recursion is as follows. Let  $\tilde{p}(m) = 0$  for  $m < 0$ ,  $\tilde{p}(0) = 1$  and

$$\tilde{p}(m) = \sum_{i=1}^P \frac{r_i}{m} \tilde{p}(m - r_i) \cdot F_i, \quad (3.20)$$

and the probability of having  $m$  units of capacity occupied is obtained after normalisation, namely:

$$p(m) = \frac{\tilde{p}(m)}{\sum_j \tilde{p}(j)}. \quad (3.21)$$

If we suppose a system with moderate load, this would mainly dwell in non-blocking subspaces (arrivals and switching to any phase is allowed) and blocking parts will have low steady state probability. In this case, the product form stationary distribution defined by (3.14) well approximates the steady state distribution. On the other hand, the local balance equations can be derived and are valid in the blocking subspaces, as it was described above. So the heuristic idea of approximating the system is: treat the states of non-blocking subspaces with the multiplier factors derived and as if these were valid in all the state space, treat the states of blocking subspaces with the multiplier factors derived from the changed service time distribution, as if it were true in the whole state space.

Considering these, the main heuristic idea is to define a modified version of (3.20) recursive formula. The modification considers the multiplier factors appearing in (3.20). Namely I use the

load dependent multiplier factors defined above. This means that when adding an element in the recursion,  $F_i$  is not constant, rather its value for channel occupancy value  $0 - r_i$  is used. This approach will not provide exact results, as this is based on product form approximations, but as we describe it in Section 3.5 the results have reasonably small error.

Thus I introduce the modified version of Kaufman-Roberts formula, applicable for approximating occupancy probabilities in the multiclass M/PH/ $C_0$  queue with phase dependent capacity requirements. The following formula is expressed with the parameters of the current application in this thesis, namely the model of a mobile radio cell.

I define  $\tilde{p}(m)$  and  $p(m)$ , the relative and the normalized probability of that  $m$  amount of capacity is occupied in equilibrium.  $\tilde{p}(m)$  is computed as  $\tilde{p}(m) = 0$  for  $m < 0$ ,  $\tilde{p}(0) = 1$ , and for  $m > 0$

$$\begin{aligned} \tilde{p}(m) &= \sum_{k=1}^K \sum_i \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{(k),N}(m - r_i^{(k)}) + \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{(k),H}(m - r_i^{(k)}) \\ p(m) &= \tilde{p}(m) \frac{1}{\sum_{m=0}^{C_0} \tilde{p}(m)}. \end{aligned} \quad (3.22)$$

This recursive formula, along with the definition of the multiplier factor (3.19) are the main results of all the modelling work and queueing system definition. Using (3.19) and (3.22) allows the determination of channel occupancy probabilities very efficiently, despite the huge state space of the model.

### 3.4.1 Performance parameters

Here I present the key session level performance indicators that allow the characterisation of the radio network. If the channel occupancy probabilities are given as (3.22), the performance parameters of the system are calculated as follows.

The call blocking probability in case of applying policy 1 for a type  $k$  call initiated in the cell is:

$$p_B^{(k),N} = \sum_{i=1}^{N_Q^{(k)}} q_i^{(k),N} \cdot \sum_{m=C_{k,N}-c_i^{(k)}+1}^{C_0} p(m). \quad (3.23)$$

The same measure for handover calls is calculated analogously.

If we denote the minimum possible capacity requirement of a type  $k$  call with  $c_{min}^{(k)}$ , the call blocking probability for a type  $k$  call initiated in the cell applying the second admission policy

has the form of:

$$\hat{p}_B^{(k),N} = \sum_{i=1}^{N_Q^{(k)}} q_i^{(k),N} \cdot \sum_{m=C_{k,N}-c_{min}^{(k)}+1}^{C_0} p(m). \quad (3.24)$$

The channel utilization is simply given as:

$$\varrho = \sum_{m=0}^{C_0} m \cdot p(m). \quad (3.25)$$

## 3.5 Numerical results: Analysis of UMTS downlink channel

The analysis presented in this Section is accomplished in order to examine the performance of 3G UMTS radio interface in the downlink direction, with the modelling framework and approximate analytical method proposed in the previous Sections. Results were compared to that of simulations.

### 3.5.1 Overview of basic UMTS radio operation

In order to understand the assumptions applied here, a brief introduction to the operation of UMTS radio interface is presented in this Section. In downlink direction the UMTS utilizes OVSF spreading codes to provide channels with different bitrates to different customers. The time is divided into frames of 10 ms length, each containing 38400 complex chip symbols, feeding a QPSK modulator. Because of the four level modulation two symbols are transmitted during a chip period, resulting in the doubling of the effective chiprate and corresponding bitrates. For every user a spreading code of the Walsh-Hadamard codes (OVSF codes) is selected. UMTS data is transmitted through dedicated physical channels, that are realized by means of a selected Walsh-Hadamard codeword for the user (theoretically more codewords might be used in parallel to define higher rate physical channels, but practical systems do not support this operation, however the HSDPA extension fundamentally contains the use of more codes creating a single high-speed channel). The complex spreading is effectively achieved by applying two OVSF codes on the real and imaginary branches of the QPSK modulation. These codes can be generated recursively using:

$$\mathbf{W}_0 = \begin{bmatrix} 1 \end{bmatrix}$$

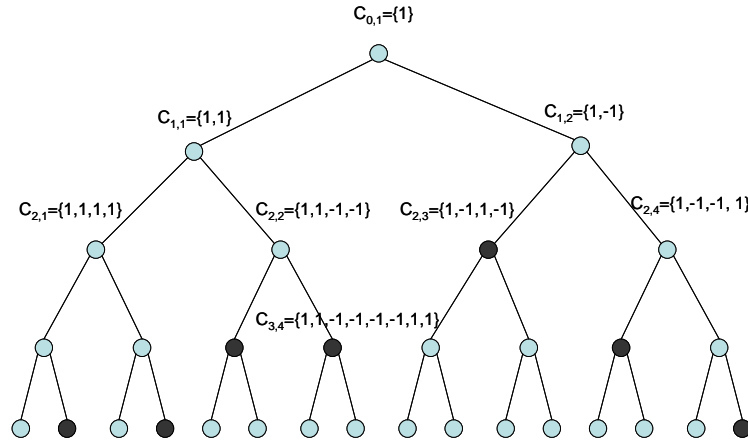


Figure 3.5: Code tree and code assignments

and

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{W}_{i-1} & -\mathbf{W}_{i-1} \\ \mathbf{W}_{i-1} & \mathbf{W}_{i-1} \end{bmatrix},$$

where each row of matrix  $\mathbf{W}_i$  contain a spreading code of length  $2^i$  chips. The maximal code length assigned to a user in UMTS downlink is 512 chips (per bit), resulting in the transmission of  $2 \cdot 75$  bits per frame, i.e. a rate of 15 kbps. The shortest applicable code is of length 4, resulting in the transmission of 19200 bits per frame, namely a rate of 1920 kbps. The set of assignable codes are often represented using a code tree, where the nodes mean codes. The tree is generated recursively similar to the code matrix: placing  $C_{0,1} = 1$  code in the root, the descendant codes of code  $j$  at level  $i$  ( $C_{i,j}$ ) – meaning one of the codes of length  $2^i$  – are  $C_{i+1,2j-1} = \{C_{i,j}; C_{i,j}\}$  at the left branch and  $C_{i+1,2j} = \{C_{i,j}; -C_{i,j}\}$  at the right branch. This tree view has the advantage of showing the so called code blocking phenomenon. Namely, not all Walsh-Hadamard codes are orthogonal, but those that are not ancestors or descendants of each other. This means that in any moment the codes allocated to users may not be arbitrary, but just orthogonal codes, to assure separation of different users' signals at the receiver. Figure 3.5 shows an example of code tree with maximum code length 16. As it is clear from the composition of codes, if we denote the maximum theoretical bitrate (achieved by using length 1 code, i.e. no spreading, only one user is serviced at once) by  $R$ , then a code with length  $2^i$  (code  $C_{i,j}$ ) results in a bitrate  $R/2^i$ . At any moment the code allocation results in the total rate of all customers being less than  $R$ , namely if  $n_i$  connections use codewords of length  $2^i$ , then  $\sum_i n_i \cdot R/2^i \leq R$ . But due to the nonorthogonality of Walsh-Hadamard codes of different lengths a code allocation may happen,

where the total rate of customers is low enough to admit one more customer with higher rate (i.e. shorter code), but this cannot happen because of the lack of free orthogonal codes. In Figure 3.5 an example is shown, with the assigned codes denoted by dark nodes. Here the total occupied rate is  $R \cdot \frac{13}{16}$ , but no more connection with code length 8 (rate  $R \cdot \frac{1}{8}$ ) can be admitted because of code blocking. Thus code blocking is the phenomenon that arises when sufficient capacity is available in a CDMA system, but connections with certain rates may not admit because of the inefficient allocation of spreading codes. Returning to the example of Figure 3.5, assigning to one of the low rate users code  $C_{4,3}$  or  $C_{4,15}$  instead of  $C_{4,2}$  would free code  $C_{3,1}$ , allowing a new connection with code length 8 to enter the system.

It is clear that optimal allocation of codes can vanish code blockings, but this requires the rearrangement of the code allocation scheme after a code is disengaged (transmission terminated). Regarding the situation presented at Figure 3.5 it is clear that when code  $C_{4,16}$  is disengaged, then reallocating  $C_{4,15}$  and  $C_{4,16}$  instead of  $C_{4,2}$  and  $C_{4,4}$  makes two codes of length 8 or one code of length 4 available. The problem of optimal assignment and rearrangement of codes was addressed in several papers, e.g. [78], [105], [106], [107], [108].

In our numerical example we suppose that optimal channelisation code assignment and reassignment take place, therefore code blocking does not happen. As described here, the space of orthogonal spreading codes can be viewed as the capacity of a UMTS carrier in a cell. However, as it will be detailed in forthcoming sections of this dissertation, this is not the only dimension of the capacity. In reality, the finite base station transmit power is the bottleneck capacity of a cell, due to interference caused by neighboring cells and self-interference that occurs in multipath environment, due to the loss of perfect orthogonality between OVVSF codes. However, for this particular set of results I consider only the code capacity of the system, interference is not taken into account, therefore the following results are dealing with the theoretical maximum performance of an UMTS carrier. Thus, these examinations consider the case when radio environment is "friendly" (flat area, users dwelling mainly in the Node B's vicinity, no multipath effect, abundant Node B power).

In practice in the UMTS downlink traffic channels and dedicated signalling channels are time-multiplexed, meaning that each frame carries a number of traffic and some signalling bits. In the following example we suppose that if a source requires a given transmission rate, that includes the necessary signalling information as well, hence the capacity requirement of a transmission is given by the necessary gross physical bitrate, containing signalling, channel coding and other overheads. Moreover, in UMTS it is not allowed to exploit the total downlink capacity by user

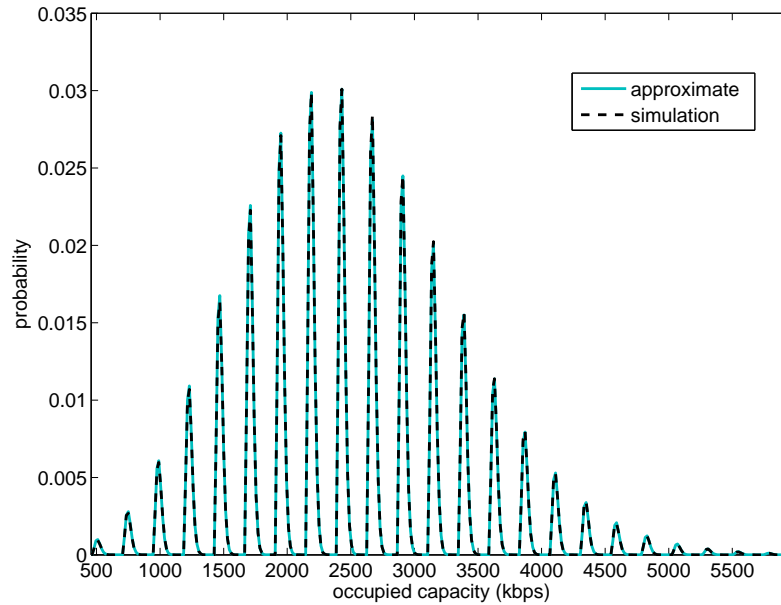


Figure 3.6: Capacity occupation in lightly loaded cell

traffic, since some codes should be preserved to create other signalling channels. Here we neglect this, so all the capacity can be used by traffic. But the effect of signalling channels can easily be included in the proposed modelling framework, namely by setting the maximum amount of capacity that connections of any class may occupy smaller than the total capacity.

### 3.5.2 Parameters used in the analysis and results

Here I present all the actual distributions, parameters and customer types that were used in the analysis as well as numerical results of actual investigations. I examined a cell with the proposed modelling framework. User mobility was described by residual dwell times and dwell times. These were modelled to follow lognormal distributions as it was suggested in [63]. Here I chose lognormal distributions with means 6.55 and 8.84 minutes and variances 18.6 and 33.9 minutes for residual dwell time and dwell time respectively. Mobility descriptors were the same for each class and the distributions were approximated by properly chosen 6 phase PH distributions.

I supposed three types of connections:

- The first type is speech, that lasts for exponentially distributed time with mean 2 minutes and generates traffic of exponentially distributed spurts of 15 kbps (corresponding to codes

of length 512) and exponentially distributed silence periods (no transmission), i.e. a classical ON-OFF model. In practice, the traditional AMR (Adaptive MultiRate) voice codec of UMTS generates speech data at the rate of 12.2 kbps maximum, requiring the use of 64 kbps physical bitrate, and the constant occupation of a corresponding channelisation code of length 128. However, in this example I supposed VoIP transmission, with a low data rate predictive voice codec with activity detection, hence the required low physical rate and the simple ON-OFF structure.

- The next type is streaming video, with normally distributed duration, approximated by a proper 8 phase PH distribution for new connections, and a PH residual session length calculated by the method described in Chapter 4. Streaming video is supposed to generate traffic at constant physical rate of 240 kbps (meaning the spreading factor is 32). This bitrate corresponds to a low resolution and low frame-rate video transmission of 96 kbps and the attached 16 kbps voice stream with channel coding.
- The third session type is interactive data session, that last for an exponentially distributed time with mean 15 minutes, generating 480 kbps (code length 16) bursts that last for a Weibull distributed time with mean around 0.5 second and OFF periods that are also Weibull distributed with mean half minute (this kind of source model was suggested in [92]), both Weibull distributions are approximated by 14 phase PHs. The 480 kbps physical bitrate corresponds to a 144 kbps data bearer service, that has robust channel coding.

The analysis an UMTS downlink channel with such connection types was carried out by the proposed approximate Kaufman-Roberts method and computer simulations. During simulations the user descriptor times were simulated according to their fitted PH distributions, therefore the results contain the inaccuracy introduced by the product form approximation, not the PH fitting. Moreover, in real life problems the adequate method of obtaining user describing distributions is to fit a PH distribution according to measured or simulated data, namely other distributions that appear in the literature are also risen as approximation of measurement data and the ability of arbitrary PHs to approximate an unknown distribution is better than that of a special distribution (e.g. Weibull, lognormal, etc.). The reason why we described the original supposed distributions is to show that the presented framework is applicable for evaluating arbitrary scenarios.

In Figure 3.6 the channel occupancy probabilities of the UMTS downlink are plotted. In this system the basic capacity unit is 15 kbps (SF= 512 codes), therefore it is the unit of the X axis. The total capacity is the theoretical maximum achievable by no coding, it is 7680 kbps

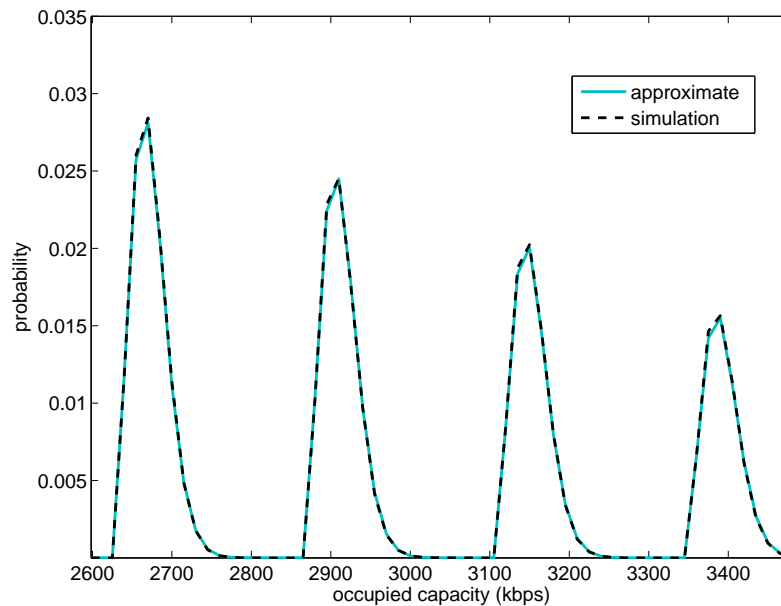


Figure 3.7: Capacity occupation in lightly loaded cell

in this case. It was supposed that each class new and handover connections may occupy all the capacity and admission control policy 1 (immediate blocking) is applied. The system was lightly loaded, resulting in a channel utilisation of 32%. Simulations were run for 50000 minutes of system time, that resulted in approximately 300000 arrivals. The system is lightly loaded, shown by channel occupancy probability disappearing after 5000 kbps. It is apparent, that the results obtained by the approximate Kaufman-Roberts method and simulations are indistinguishables. The special shape of the curve is due to the fact that multiples of 240 kbps occupancies have the highest probabilities, as this is the rate of streaming video connections, as well as half of the rate of interactive data bursts. To get a closer view, on Figure 3.7 a section of the previous curve is plotted. The results obtained by the two methods are almost identical, as it was anticipated in a lightly loaded system. A system with heavier load, resulting in a 72% utilisation is also evaluated. Channel occupancy probabilities are presented on Figure 3.8. The simulations were run for 17650 minutes of system time, resulting again in approximately 300000 arrivals. We can see that under heavier load conditions the approximate method works with less accuracy; we anticipated this since the probability of blocking sub-spaces increased in this case. However, this accuracy is still reasonable and results in just a slight difference in performance parameters. It is apparent in Figures 3.6-3.8 that the channel occupancy probabilities are plotted with solid

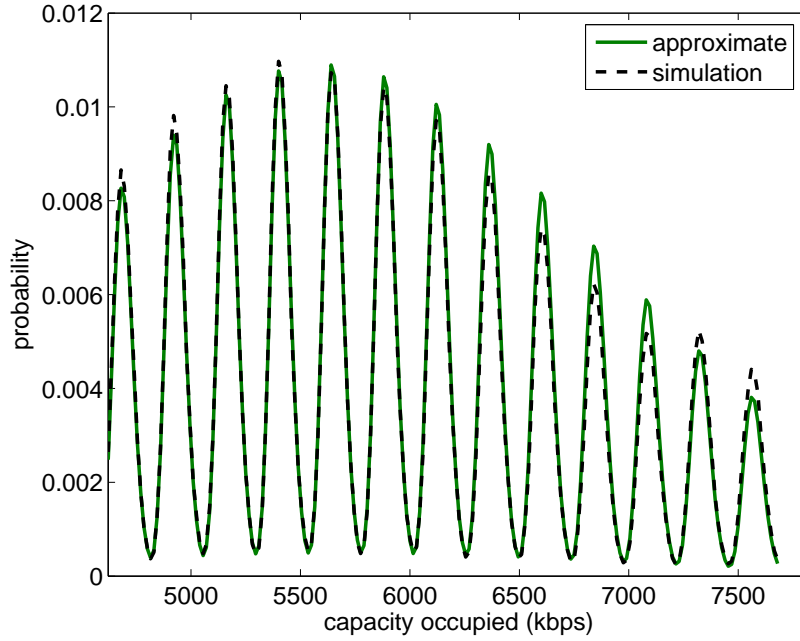


Figure 3.8: Capacity occupation in loaded cell

curves, despite the discrete possible values of total occupied capacity. The solid plot was chosen for better visibility purposes. The plots are created using 15 kbps granularity of the capacity axis (because this is the smallest unit the occupied capacity may change with). Thus, the reason of zero occupancy probabilities is that especially in lightly loaded cases the probability of occupying that amount of capacity is close to zero. Naturally, Figure 3.8 showed an example when load is heavier, here occupancy probabilities do not reach zero (for capacity values shown in the Figure).

To get a deeper insight of the accuracy of the approximate method we created Figure 3.9, where the deviation of the approximate results from that of simulation is plotted as the system load increases. At the original configuration (Figure 3.6) the incoming rate of all classes were 1 per minute. As accuracy measure, the left side of Figure 3.9 plots the sum of the absolute values of the differences of channel occupancy probabilities calculated by the two methods, i.e. the measure of accuracy is  $m = \sum_i |p_{sim}(i) - p_{K-R}(i)|$ , where  $p_{sim}(i)$  is the probability of having  $i$  amount of capacity occupied, obtained by simulation and  $p_{K-R}(i)$  is the same quantity calculated by the approximate method. Note that this accuracy measure is simply the sum of the deviation of the results. The accuracy measure is plotted as the rate of new and handover data sessions grow while the rate of other connections remains the same, similarly accuracy is plotted as the rate

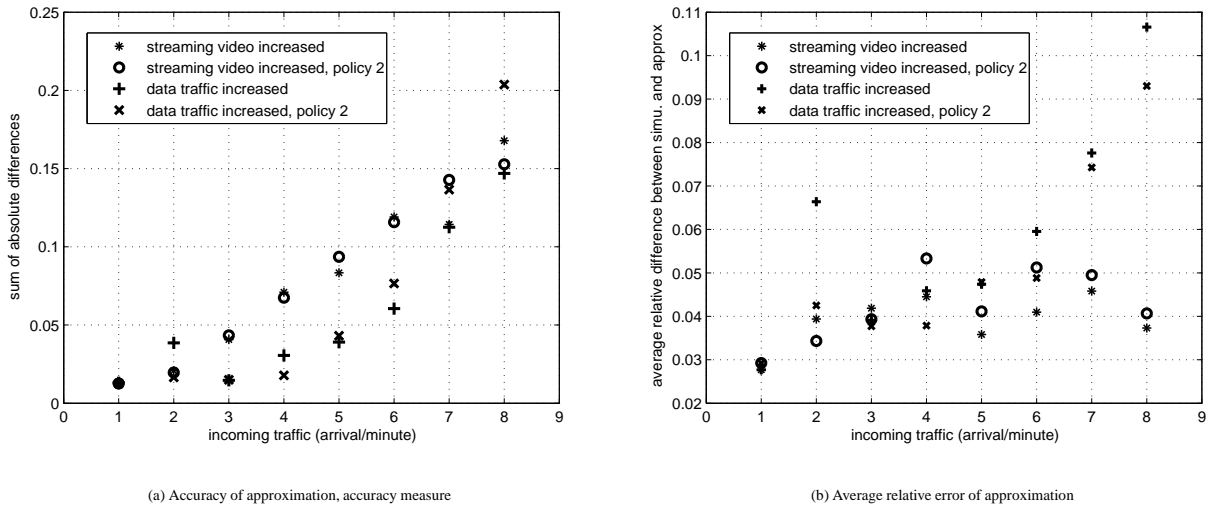


Figure 3.9: Accuracy of the approximation

of streaming video connections increase, while all others do not change. The case of applying admission control policy 2 (i.e. rate reduction) on interactive data sessions (both handover and newly initiated) is also plotted. On the X axis the incoming rate of both handover and new connections is represented in all cases, so the arrival rate was increased from 1 to 7 arrivals per minute. The results show, that the sum of deviations of occupancy probabilities are around 0.15 under the highest load conditions, knowing that this result rises from summing up 513 differences indicates the accuracy of the approximate algorithm. The right hand side of Figure 3.9 plots the average value of the relative absolute differences, that is  $m = \frac{1}{I} \sum_i \frac{|p_{sim}(i) - p_{K-R}(i)|}{p_{sim}(i)}$ , where  $I$  denotes the total number of possible occupancy values (513 in the particular examples). As it is apparent, when the load is increased by means of the constant bitrate streaming video connections, the average relative accuracy stays around 0.04 – 0.05, because in this case the state space would rather dominated by the fixed rate connections, making it similar to a product form space. On the other hand, when the ON-OFF type data traffic generates the load, the average deviation of the results increase, as it was anticipated. However, for very high loads this measure is still around 0.1.

Figure 3.10 shows the utilisation of the downlink UMTS channel as system load increases. The values obtained when streaming load is increased are practically the same for the approximate method and simulations, either policy 1 or policy 2 is applied on interactive data. This is because the state space is governed by the arrival of constant rate streaming connections, making

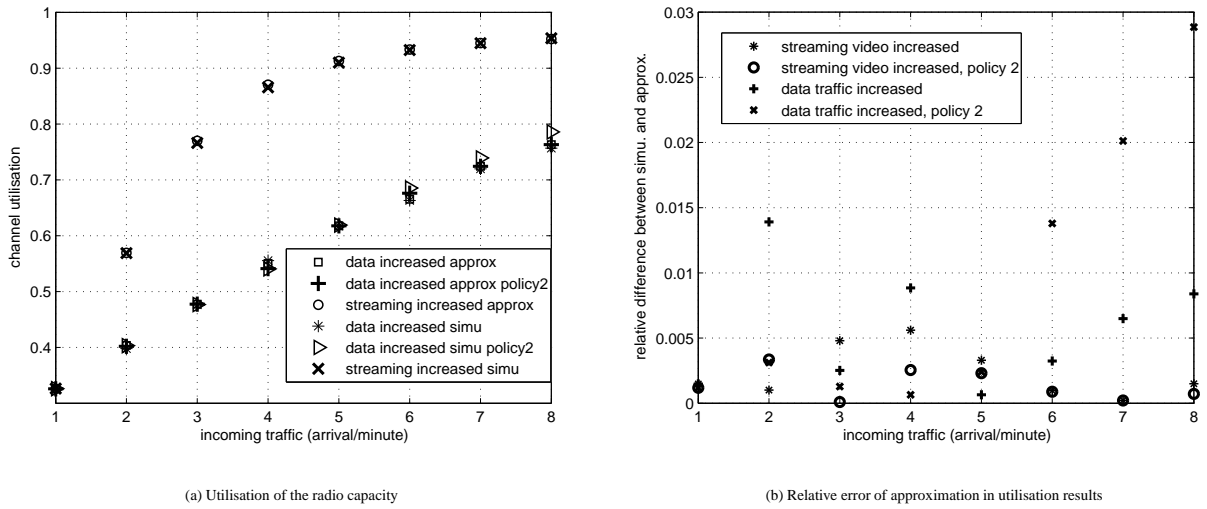
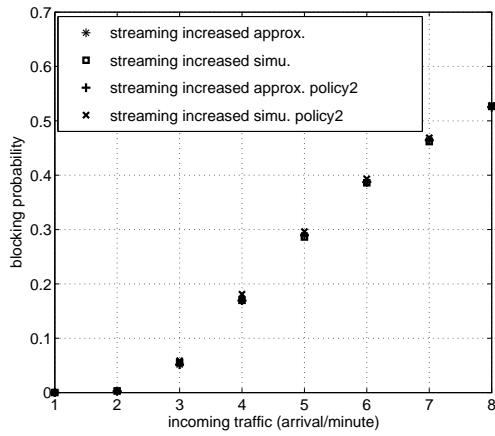


Figure 3.10: Utilisation of the channel (left) and accuracy of approximation (right)

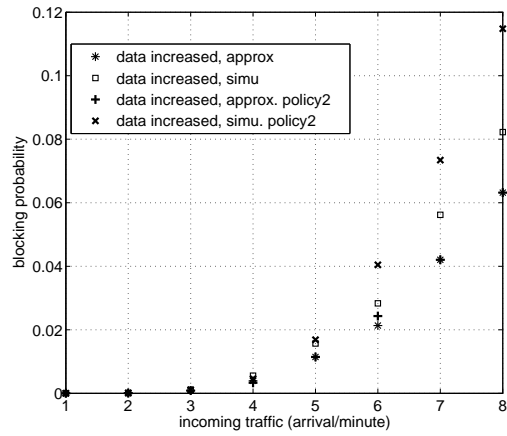
the problem similar to a product form one. Thus for this case only the result of the immediate blocking policy is plotted. In the case when the rate of data sessions is increased, the results are slightly different for the two policies and depend on the means of calculation (i.e. simulation or the approximate method). The right plot in Figure 3.10 shows the relative difference of the utilisations, calculated by the proposed approximation and simulation. Formally, what is plotted is the amount  $\frac{|U_{sim} - U_{K-R}|}{U_{K-R}}$ , where  $U_{sim}$  is the utilisation value gained by means of simulations, whereas  $U_{K-R}$  is the utilisation obtained based on the proposed Kaufman-Roberts based approximation. Apparently the difference is practically zero when streaming load is increasing, but in case of high data loads the utilisation values differ by less than 3 percent.

Figure 3.11 gives insight of the performance parameters of the UMTS downlink channel with the described connection types. Both cases of increasing handoff and new connection arrival rates of either streaming video or interactive data sessions is examined. The former case is plotted on the graphs at the top of Figure 3.11, while the latter is at the bottom. Regarding the blocking performance of streaming connections, we may conclude that since no capacity reservation was applied, the blocking probabilities of handoff and newly initiated sessions are equal.

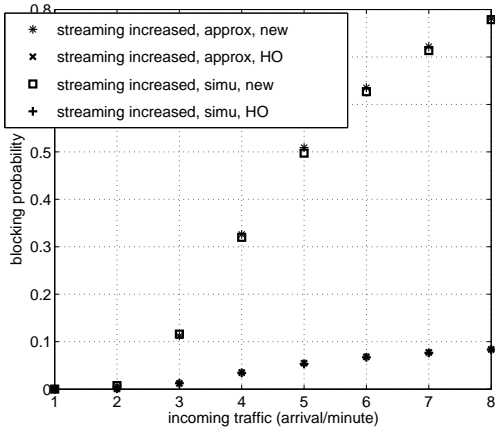
The effect of applying admission policy 2 on interactive data connections is also evaluated. Regarding data sessions, the blocking probabilities of new and handover connections are different, since according to the applied model a new session always begins with an ON burst, while for handoff connections the burst process is supposed to have reached equilibrium until the in-



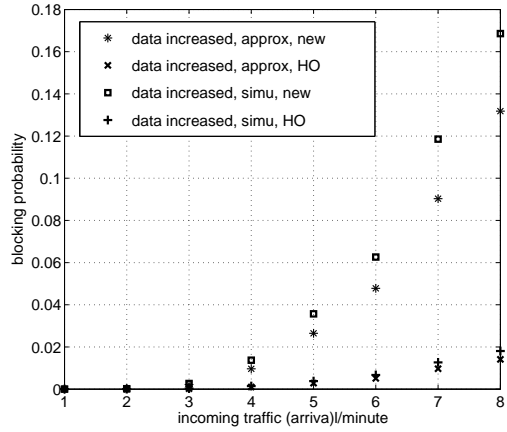
(a) Blocking probability of streaming video



(b) Blocking probability of streaming video



(c) Blocking probability of interactive data



(d) Blocking probability of interactive data

Figure 3.11: Blocking probability of streaming (top) and data (bottom) connections

stant of handoff, thus the connection is in OFF state with higher probability so it is not blocked. In this case the application of policy 2 on data sources is not evaluated, since according to the definition of the policy this would result always in zero blocking.

One may observe the following: if the system is loaded by increasing the streaming video traffic, the approximate method gives very good results even under extremely heavy loads. This is because in this case the majority of the capacity is occupied by the constant rate streaming connections and the state transitions are mainly determined by the arrival and leaving rates of these. So we intuitively feel that in this case the approximating the system with a product form solution is more accurate, since having only constant rate applications would result in a product form solution, thus the modified Kaufman-Roberts method would provide exact results.

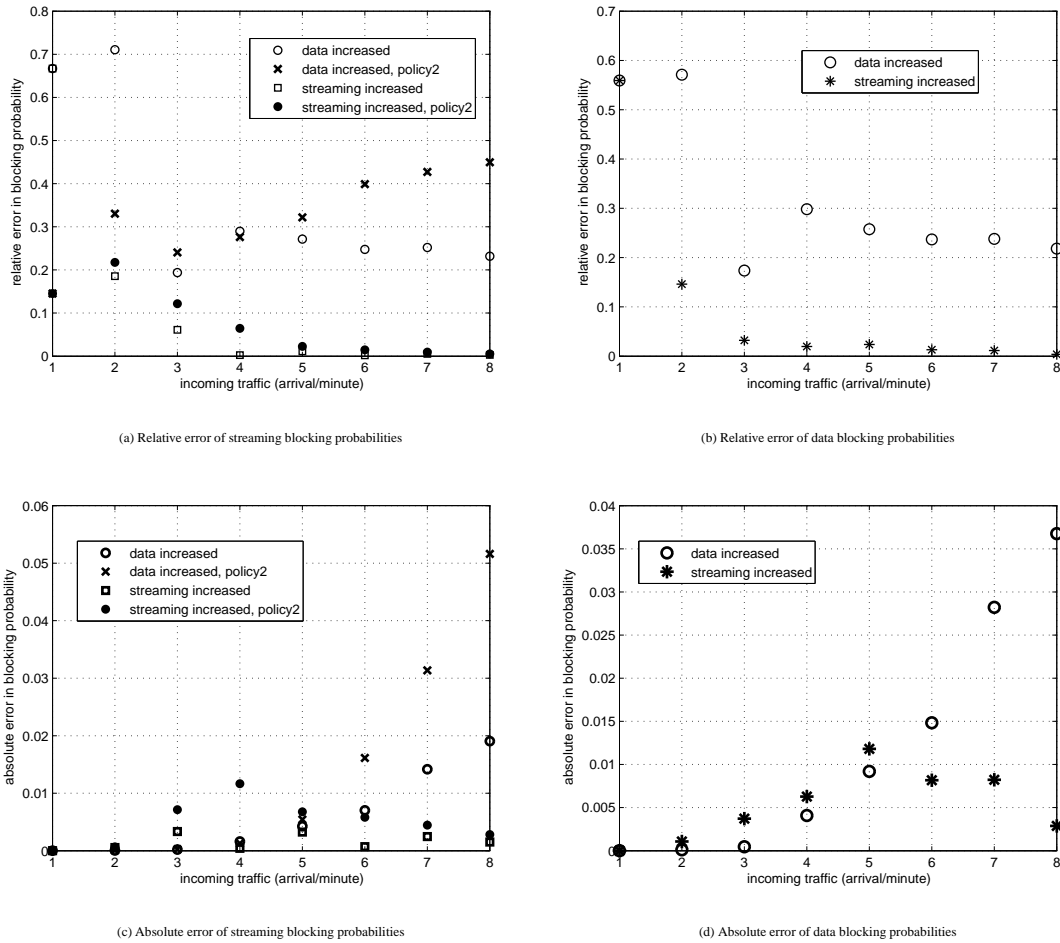


Figure 3.12: Relative (top) and absolute (bottom) error of the approximation in blocking probabilities of streaming (left) and data (right) connections

More inaccuracy is introduced when the volume of data traffic causes the cell to be heavily loaded, since in this case the state space is mainly ruled by the transitions of the general ON-OFF traffic. Since the system dwells in states near the border of the state space with higher probability, the product form approximation becomes less accurate. But we may conclude, that in a reasonable network the blocking probability of a streaming video connection should not exceed 0.02, data connections may acceptably be blocked with a bit higher probability, say around 0.04 and under load conditions resulting in these blocking measures the approximation still provides acceptable accuracy.

In the top of Figure 3.12 the relative error of the blocking probabilities calculated by the approximation and simulations is shown (again, this is the absolute value of the difference be-

tween the blocking probabilities obtained by simulations and the approximation, divided by the result of simulation). The left hand side shows this relative error for the blocking probabilities of streaming, whereas the right hand side for data connections. We might observe quite high values approaching 70 percent, contraintuitively for the case of lower loads. The reason of this is the following. For very light loads, the value of blocking probability is close to zero. In this case a minor deviance, in absolute value, will result to be high as relative value. As example, if the simulation results in 0.0001 and the approximation in 0.0002, the relative error is 100 percent, however the difference in actual values is negligible. Therefore we might conclude, that the relative deviations in terms of blocking probabilities might not be as informative.

For this reason, the bottom of Figure 3.12 shows the absolute errors in the blocking probabilities. One might observe, that the absolute error can be as high as 0.02 – 0.05 for higher loads, that might seem unacceptably high, as this is the allowed range for blocking probabilities in a real network. However, comparing this to the blocking probabilities itself shown in Figure 3.11, we can conclude that this high inaccuracy occurs only when the system is so overloaded, that blocking probability well exceeds 0.1, meaning that presumably no system would be designed to operate under such heavy traffic circumstances. As conclusion, under realistic traffic loads the proposed approximation gives very good results.

{

## Chapter 4

# Calculating the residual session length distribution

In Chapter 2 the meaning of residual session length or residual connection holding time was introduced. It is not straightforward to determine the distribution of this time. Hence this Chapter is devoted to the investigation of the residual session length distribution of connections attaching to a particular base station via handover. To calculate this, the distribution of the session length, the dwell time and the residual dwell time is needed, along with topological information on the cells of the area surrounding a particular examined base station. If the residual connection holding time distribution is derived, the blocking performance and utilization of the cell can be calculated, for instance by the method we described in the previous Chapter. The problem analysed in this Chapter is very weakly covered in the literature. Among publications of the recent years [109] could be seen as dealing with the similar problem, but with covering a much narrower scope.

It has to be noted that for practical usability purposes (in terms of using the modelling framework described in previous Chapters) the residual session length distribution might be approximated with a simpler distribution, having the same mean and variance as of the actual residual session length, as would not greatly influence the system level performance results achieved by the model. However, in order to sketch the modelling framework in full details, we see the determination of the actual distribution of the residual connection duration as an interesting meaningful problem, that actually can be solved.

## 4.1 Modelling assumptions

In this Section I present the cellular network in which the following calculations targeting the residual session length distribution are carried out.

The aim is to determine the residual session length distribution of handover customers arriving to a particular base station in a cellular system. The cellular structure is supposed to have the following properties. It consists of several homogeneous cells, meaning that the dwell time and residual dwell time distributions are identical in all cells of the examined area. This proposition models the case when the network covers an area with mostly identical or similar radio environments and with base stations transmitting with equal powers, hence the cell sizes are equal, moreover the users' mobility patterns and speeds are identical from the resulting dwell time point of view. However, the generalisation of the following calculation with different dwell time distributions is possible.

Naturally, the dwell time depends on not only the cells but the type of users roaming throughout the region (e.g. pedestrians, slow vehicles, fast vehicles, etc.). In case of several customer types in terms of mobility, the following analysis should be carried out for each type independently, since we propose that a customer does not change its mobility class during a connection. The following paragraphs suppose a single session length distribution as well, although this parameter may also depend on the connection class. Assuming that the connection type of a customer does not change during transmission, the presented calculation should again be performed for all classes independently.

We suppose that for a given mobility and connection class there is a maximum number of  $I$ , so that the probability that a customer performs more than  $I$  handovers is negligible. Therefore the residual session length is influenced only by those cells that can be reached by less than  $I + 1$  handovers from the examined cell. There is no restriction on the shape and nearness of the cells.

The mobility of the users was described by the dwell times within a cell, but for the final goal the movement between cells must also be characterized. We assume that while roaming, a mobile enters cell  $l$  after leaving cell  $k$  with probability  $\Pi_{kl}$ , regardless of its previous route among cells. These values are collected in routing matrix  $\Pi$ . Obviously those entries of this matrix that correspond to not neighboring cells is 0.

One more parameter is needed, that is the distribution of connection initiation among cells. Namely, a probability vector  $\underline{B}$  is given to describe session initiation density in the region, with  $B_i$  meaning the probability that if a new session is initiated in the area that happens in cell  $i$ .

Customers of a particular mobility and connection class are characterized by their session length distribution. Its distribution function and probability density function is denoted by  $F(x)$  and  $f(x)$  respectively, the same descriptors of the dwell times and residual dwell times are  $G(x)$ ,  $g(x)$ ,  $G_R(x)$  and  $g_R(x)$ .

## 4.2 Calculating the Residual Session Length

This Section presents the calculation method for the residual session length distribution, for general distributions and for phase type distributions as well.

### 4.2.1 General Distributions

Let us consider the event that a mobile customer attaches to the examined base station after completing its  $i$ th handover. This means that a residual dwell time plus  $i - 1$  dwell times have elapsed after the instant of connection initiation. Let this time be denoted by  $\tau^{(i)}$  and its density function by  $g^{(i)}(x)$ . In this case the residual session length is  $\tau_{S,R}^{(i)} = \tau_S - \tau^{(i)}$ , where  $\tau_S$  is the session length and  $\tau_{S,R}^{(i)}$  is the residual session length conditioned on the case when the connection arrives after the  $i$ th handover. First, we are interested in the distribution function of  $\tau_{S,R}^{(i)}$ , that is

$$F_R^{(i)}(t) = \Pr(\tau_{S,R}^{(i)} < t) = \Pr(\tau_S - \tau^{(i)} < t | \tau_S > \tau^{(i)}) = \frac{\Pr(\tau^{(i)} < \tau_S < t + \tau^{(i)})}{\Pr(\tau_S > \tau^{(i)})}. \quad (4.1)$$

After the expression of probabilities in (4.1), we arrive to the distribution function of the residual session length supposing that the connection arrives via its  $i$ th handover is

$$F_R^{(i)}(t) = \frac{\int_0^\infty (F(t+x) - F(x))g^{(i)}(x)dx}{\int_0^\infty (1 - F(x))g^{(i)}(x)dx}. \quad (4.2)$$

The next goal is to derive the probability of the event that an incoming handover is the  $i$ th cell change of the connection. If this quantity denoted by  $p^{(i)}$  is given, then the resultant residual session length has the distribution function of

$$F_R(t) = \sum_{i=1}^I p^{(i)} \cdot F_R^{(i)}(t). \quad (4.3)$$

Assuming general dwell time distributions one may have two options to calculate  $g^{(i)}(x)$  of equation (4.2). One is to perform multiple convolutions of the pdfs of the residual dwell time

and  $i - 1$  dwell times, i.e.

$$g^{(i)}(x) = g_R(x) * g(x) * \dots * g(x). \quad (4.4)$$

The other option is to determine the Laplace transforms of the pdfs of residual dwell times and dwell times (denoted by  $G_R(s)$  and  $G(s)$  respectively), multiply them and inverse Laplace transform the result:

$$g^{(i)}(x) = \mathcal{L}^{-1} \{G_R(s) \cdot G(s)^{i-1}\}. \quad (4.5)$$

In most cases both approaches and therefore the calculation of (4.2) can only be performed numerically, thus all  $F_R^{(i)}(t)$ -s of (4.3) will be available as series of numerical data. Nevertheless one would need the distribution function in analytical form to perform further system analysis. The solution may be to fit some distribution to the resultant numerical data and use this. Moreover, convolutions and numerical Laplace transforms are rather time consuming tasks, requiring huge amount of computational capacity.

## 4.2.2 Phase Type Models

By applying the family of Phase Type distributions as models of the dwell time and session length distribution, we may overcome some drawbacks of the previous calculations and the distribution function of the residual session length will be available in analytical form and ready to use in analytical system models.

Let us denote the parameters of the PH distributed session length by  $\underline{l}$ ,  $\underline{L}$  and  $\underline{L}^0$ . Substituting the distribution function from (2.7) with these parameters into (4.2) we get:

$$F_R^{(i)}(t) = \frac{\int_0^\infty \underline{l} \cdot e^{\underline{L}x} \cdot \underline{h} \cdot g^{(i)}(x) dx - \int_0^\infty \underline{l} e^{\underline{L}(t+x)} \cdot \underline{h} \cdot g^{(i)}(x) dx}{\int_0^\infty \underline{l} \cdot e^{\underline{L}x} \cdot \underline{h} \cdot g^{(i)}(x) dx} = 1 - \frac{\int_0^\infty \underline{l} \cdot e^{\underline{L}x} \cdot g^{(i)}(x) dx \cdot e^{\underline{L}t} \cdot \underline{h}}{\int_0^\infty \underline{l} \cdot e^{\underline{L}x} \cdot \underline{h} \cdot g^{(i)}(x) dx}. \quad (4.6)$$

By comparing this with the distribution function of the PH (2.7) we may observe that the resultant distribution is also PH, with the same phase structure as the original session length has, only its initial probability vector changes. If we denote the initial probability vector of the residual session length distribution of a connection given that it arrived after its  $i$ th handover by  $\underline{l}_R^{(i)}$  we get

$$\underline{l}_R^{(i)} = \frac{\int_0^\infty \underline{l} \cdot e^{\underline{L}x} \cdot g^{(i)}(x) dx}{\int_0^\infty \underline{l} \cdot e^{\underline{L}x} \cdot \underline{h} \cdot g^{(i)}(x) dx}. \quad (4.7)$$

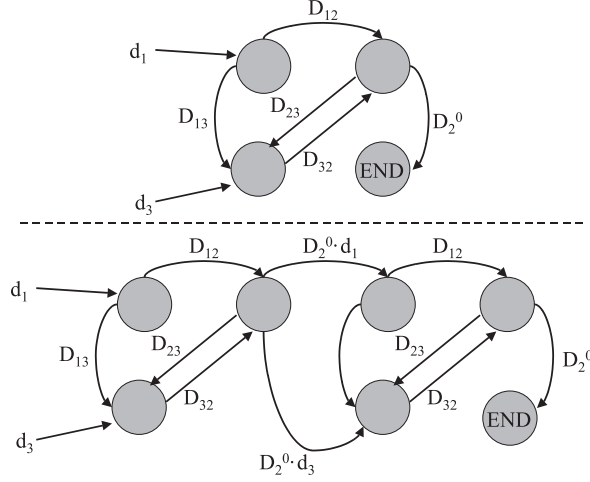


Figure 4.1: Composition of sum of PH distributions

It straightly follows, that in this case the resultant residual session length differs from the connection lifetime only by means of its initial probability vector, and instead of (4.3) one has to calculate

$$\underline{l}_R = \sum_{i=1}^I p^{(i)} \cdot \underline{l}_R^{(i)}, \quad (4.8)$$

where again  $p^{(i)}$  denotes the probability that a customer arrives after its  $i$ th handover. Now we see that after performing these calculations the PH residual session length is known in analytical form, with parameters  $\underline{l}_R$ ,  $\mathbf{L}$  and  $\underline{L}^0$ .

Modelling the dwell time and residual dwell time by PH distributions also makes the calculation of  $g^{(i)}(x)$  of (4.2) easier. Namely, the sum of two PH distributions is also PH and can be composed as follows: the absorbing state of the first PH distribution is replaced by the phases of the one we want to add. From each phase  $i$  of the first PH into each phase  $j$  of the added PH the transmission rate is  $D_i^0 \cdot d_j$ . Figure 4.1 represents an example of adding two identical PHs with 3 phases. We denote the parameters of the PH distributed dwell time by  $\underline{d}$ ,  $\mathbf{D}$  and  $\underline{D}^0$ , that of the residual dwell time by  $\underline{d}_R$ ,  $\mathbf{D}_R$  and  $\underline{D}_R^0$ , hence the former denotation in the Figure. Considering this method of adding up PH distributions, the sum of a residual dwell time and  $i - 1$  dwell times (that has the density function  $g^{(i)}(x)$ ) is also PH distributed. Let us denote the number of phases of the dwell time by  $N_D$ , that of the residual dwell time by  $N_{D,R}$ , then the summed distribution obviously has  $N_{D,R} + (i - 1)N_D$  phases. The initial probability vector of this distribution is  $\underline{d}^{(i)} = [\underline{d}_R \ \mathbf{0}]$ , where  $\mathbf{0}$  is a row vector containing  $(i - 1)N_D$  zeros, its rate matrix  $\mathbf{D}^{(i)}$  is a

hypermatrix with the following structure:

$$\mathbf{D}^{(i)} = \begin{bmatrix} \mathbf{D}_R & \underline{D}_R^0 \cdot \underline{d} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \underline{D}^0 \cdot \underline{d} & \mathbf{0} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{D} & \underline{D}^0 \cdot \underline{d} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{D} \end{bmatrix}.$$

The column vector containing the rates into the absorbing state of this summed PH has  $N_{D,R} + (i - 2)N_D$  zeros and the elements of  $\underline{D}^0$  at the last  $N_D$  entries. When calculating (4.2),  $g^{(i)}(x)$  has the form of (2.7) with these parameters.

### 4.2.3 Determining handover probabilities

To calculate the quantity  $p^{(i)}$  of (4.3) and (4.8) we have to consider how a user can arrive to the examined cell (without the loss of generality this cell is denoted by number 1 from now on) after having initiated  $i$  handovers. Let us denote the minimum number of handovers that is necessary to reach cell  $l$  from cell  $k$  by  $H(k, l)$ , this can be viewed as the distance of cell  $k$  and cell  $l$ . It is clear that only those customers may arrive to the examined cell 1 after the  $i$ th handover that initiated their connections in cell  $k$  so that  $H(k, 1) \leq i$ . The series of cell numbers a customer follows during its roaming to cell 1 is referred in this document as a route. A route of length  $i$  ( $i$  handovers) from cell  $k$  to cell 1 is stored in a vector of length  $i + 1$  and is denoted by  $\underline{r}^{(i)}(k)$ . For instance, a route of length 5 from cell 4 to cell 1 may be  $\underline{r}^{(5)}(4) = [4, 3, 2, 5, 2, 1]$  as it is depicted in Figure 4.2. Obviously follows from the previously described mobility model of customers that a particular cell may appear several times in a route, moreover the target cell may appear in the route not only at the last position. The probability that a mobile follows a route is the product of the corresponding elements of the routing matrix  $\mathbf{\Pi}$  described above:

$$\Pr(\underline{r}^{(i)}(k)) = \prod_{j=1}^i \Pi_{r_j^{(i)}(k)r_{j+1}^{(i)}(k)}. \quad (4.9)$$

Using these network information, we are interested in the probability that when a handover connection arrives to cell 1, it was the connection's  $i$ th cell change, this quantity was previously denoted as  $p^{(i)}$ . To determine this, we should note that since this system described is ergodic, thus:

$$p^{(i)} = \frac{N_{HO}^{(i)}}{N_{HO}}, \quad (4.10)$$

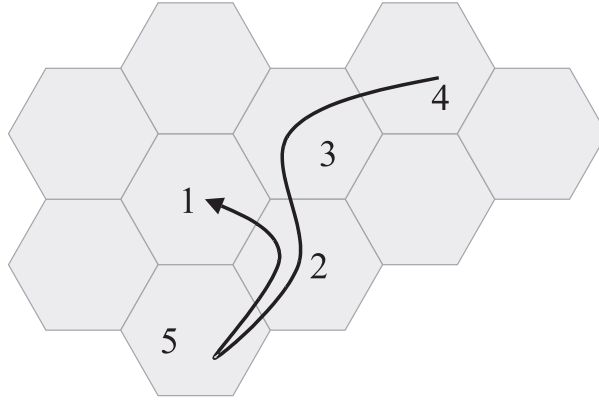


Figure 4.2: An 5 handover route to cell 1

where  $N_{\text{HO}}^{(i)}$  denote the number of connections that arrive to the reference cell after the  $i$ th handover and  $N_{\text{HO}}$  is the total number of handover connections. Let us follow what happens if a large number of  $N$  connections are initiated in the system. It is clear that  $N \cdot B_k \cdot \Pr(\tau_S > \tau^{(i)})$  of them will be initiated in cell  $k$  and last long enough to initiate at least  $i$  handovers (the last term in this product is the probability that the session duration is longer than the sum of a residual dwell time and  $i - 1$  dwell times). To determine the number of those, that arrive to the reference cell from cell  $k$  following any route of length  $i$  (denoted by  $N_{k,i}$ ), this expression should be multiplied by the sum of probabilities of all length  $i$  routes from cell  $k$  to 1. That is:

$$N_{k,i} = N \cdot B_k \cdot \Pr(\tau_S > \tau^{(i)}) \cdot \sum_{\text{all } \underline{r}^{(i)}(k)} \Pr(\underline{r}^{(i)}(k)) \quad (4.11)$$

As last step, these quantities should be summed up for all possible cells  $k$ , namely those cells that are closer than distance  $i$  to the reference cell:

$$N_{\text{HO}}^{(i)} = N \cdot \sum_{k: H(k,1) \leq i} B_k \cdot \Pr(\tau_S > \tau^{(i)}) \cdot \sum_{\text{all } \underline{r}^{(i)}(k)} \Pr(\underline{r}^{(i)}(k)) = N \cdot \hat{p}^{(i)}, \quad (4.12)$$

where  $\hat{p}^{(i)}$  is the simple denotation of the fraction of connections that we are interested in. The total number of handover connections will be simply the sum of (4.12) for all possible  $i$ s, namely as we supposed that initiating more than  $I$  handovers has negligible probability, the sum is for all  $i \leq I$ :

$$N_{\text{HO}} = N \cdot \sum_{i=1}^I \sum_{k: H(k,1) \leq i} B_k \cdot \Pr(\tau_S > \tau^{(i)}) \cdot \sum_{\text{all } \underline{r}^{(i)}(k)} \Pr(\underline{r}^{(i)}(k)) = N \cdot \sum_{i=1}^I \hat{p}^{(i)} \quad (4.13)$$

As consequence, the probability in question is:

$$p^{(i)} = \frac{N_{\text{HO}}^{(i)}}{N_{\text{HO}}} = \frac{\sum_{k:H(k,1)\leq i} B_k \cdot \Pr(\tau_S > \tau^{(i)}) \cdot \sum_{\text{all } \underline{r}^{(i)}(k)} \Pr(\underline{r}^{(i)}(k))}{\sum_{j=1}^I \sum_{k:H(k,1)\leq j} B_k \cdot \Pr(\tau_S > \tau^{(j)}) \cdot \sum_{\text{all } \underline{r}^{(j)}(k)} \Pr(\underline{r}^{(j)}(k))} = \frac{\hat{p}^{(i)}}{\sum_{j=1}^I \hat{p}^{(j)}}. \quad (4.14)$$

To calculate this, along with network topological and mobility description parameters we need:

$$\Pr(\tau_S > \tau^{(i)}) = \int_0^\infty (1 - F(x))g^{(i)}(x)dx, \quad (4.15)$$

that is calculated using the results of the previous section. Now given every necessary expressions, the residual session length is calculated according to (4.3) or (4.8).

## 4.3 Special scenarios

Before showing some calculated residual session length distributions, I investigate two special network layouts.

### 4.3.1 Motorway model

The first is a linear series of radio cells. This layout is the model of the part of a radio network that is covering a motorway. Customers are moving in either direction, but do not change their heading during movement. This means that from each cell  $k$  there is only one route to cell 1 and that is of length  $H(k, 1)$ . We suppose that mobiles initiate new connections evenly in each cell, that is  $B_k = \frac{1}{2I+1} = B$ . We assume that the traffic is not balanced, namely mass movement of mobiles may point to one of the two possible directions. For us this means that a customer that initiates new connection travels in one direction with probability  $P_0$ . In this case (4.12) gets a much simpler form, since a user may arrive to cell 1 after  $i$  handover only if it initiated its connection in either of the two cells that are exactly in distance  $i$  from cell 1 and the route probability is simply  $P_0$  or  $1 - P_0$ . Thus  $\hat{p}^{(i)}$  from (4.12) has the form:

$$\hat{p}^{(i)} = B \cdot P_0 \cdot \Pr(\tau_S > \tau^{(i)}) + B \cdot (1 - P_0) \cdot \Pr(\tau_S > \tau^{(i)}) = B \cdot \Pr(\tau_S > \tau^{(i)}) \quad (4.16)$$

for every  $i \leq I$ . Substituting this into (4.14) we get:

$$p^{(i)} = \frac{\Pr(\tau_S > \tau^{(i)})}{\sum_{j=1}^I \Pr(\tau_S > \tau^{(j)})}. \quad (4.17)$$

### 4.3.2 Homogeneous hexagonal layout

The other investigated scenario is an area covered by homogeneous hexagonal cells. This means that new connections are evenly distributed in the area, so  $B_k = B$  for every cell. In terms of mobility, the homogeneity means that from each cell the mobile might move to any other of the six neighbors with equal probability, that is  $\Pi_{kl} = \frac{1}{6}$  for each neighboring cells  $k$  and  $l$ .

In this case all routes of length  $i$  have probability  $\left(\frac{1}{6}\right)^i$ . When calculating (4.12) the total number of all the routes of length  $i$  starting from the cells  $k$  and ending at cell 1 is needed. Clearly, these routes might begin in a cell  $k$  that is within the  $i$  distance vicinity of cell 1 (i.e.  $H(k, 1) \leq i$ ). To determine the number of these, a simple change of directions help. Namely the set of such routes is the same as the set of all routes of length  $i$  that are initiated *from* cell 1. Since in all steps 6 possible destinations are available, the total number of such routes is  $6^i$ , each with route probability of  $\left(\frac{1}{6}\right)^i$ .

Therefore (4.12) gets the simple form of

$$\hat{p}^{(i)} = B \cdot Pr(\tau_S > \tau^{(i)}) \cdot \sum_{k: H(k,1) \leq i} \sum_{all \underline{r}^{(i)}(k)} Pr(\underline{r}^{(i)}(k)) = B \cdot Pr(\tau_S > \tau^{(i)}) \cdot 6^i \left(\frac{1}{6}\right)^i = B \cdot Pr(\tau_S > \tau^{(i)}). \quad (4.18)$$

Substituting this into (4.14), we arrive to 4.17 that is we have the same result as with the motorway model case! We may conclude that in this two unique scenarios the residual session length distribution is equal and is affected by the customer describing times only, because of the special cellular topologies.

## 4.4 Numerical results

The method was tested in an environment containing standard hexagonal cells that were homogeneous in terms of the dwell time and residual dwell time distribution of users roaming in any of the cells. The residual dwell times and dwell times were modeled to follow lognormal distributions as it was suggested in [63]. Here we chose lognormal distributions with means 6.55 and 8.84 minutes and variances 18.6 and 33.9 minutes for residual dwell time and dwell time respectively. The connection holding time was supposed to follow normal distribution with mean 7 variance 2 minutes. In this example this was supposed to model length of short streaming video connections, e.g. trailers, video clips or commercials. With these parameters the probability of initiating more than five handovers (i.e. the probability of a random normally distributed variable

being greater than the sum of five lognormally distributed variables with the above parameters) was negligible. For this reason the area under investigation contained 91 cells (cell 1 in the middle and surrounding cells not further than 5). The routing probabilities among cells ( $\Pi_{kl}$ ), as well as the connection initiation probabilities in cells ( $B_k$ ) were chosen randomly, but strictly not uniformly. The resultant routing matrix ( $\Pi$ ) was used as the connectivity matrix of the graph description of the area, with the nodes representing the cells and the edges representing the borders of neighboring cells. An auxiliary algorithm was applied to find all the routes to the given cell with length less than six, and their probabilities.

The distribution of the residual of the normally distributed connections was computed in this scenario. The results on one hand were obtained by calculating  $g^{(i)}(x)$  of equation (4.2) as multiple convolutions. On the other hand 6-phase PH distributions were fitted to the original lognormal dwell time distributions and a 8-phase PH to the normal session length, with the EM algorithm [44]. The pdfs of the original and fitted distributions are plotted in Figure 4.3. It is visible, that the PH fitting of the lognormal distribution resulted in a pdf that is hardly distinguishable from the original one, while the normal distribution was not approximated very closely by the 8 phase PH. The latter approximation naturally could be improved by fitting PH with more phases, or letting the EM algorithm to run for longer time. However, for demonstrating the validity of the calculation this approximation accuracy was enough and resulted in just slight degradation in the accuracy of the derived residual session length distribution, as it is shown in subsequent paragraphs. Figure 4.4 plots the probability density function of the residual session length in the described example system. Besides the graph obtained by the convolutional method, the effect of utilizing easier summation of PH random variables is also visible. The curve labelled by “PH fitting all” plots the density function of the residual session length if all the three descriptor distributions were replaced by a fitted PH, as mentioned above and the result was calculated according to (4.7) and (4.8). The curve labelled “PH fitting dwell” presents the result of the approach when only dwell times were approximated by PH (to avoid time consuming numerical convolutions) and the session length was taken into account with its original pdf. The Figure presents probability density function, while the method described in Chapter 4 computes cumulative distribution function. Naturally, the latter was computed, than the pdf was calculated by numerical derivation. It is apparent that both PH fitting approaches result in a pdf that slightly differs from that of obtained by convolutions, especially around time zero. As it was intuitively anticipated, substituting all three descriptor distributions with fitted PHs resulted in the most inaccurate result, because of the inaccuracy introduced by PH fitting. However, with this method

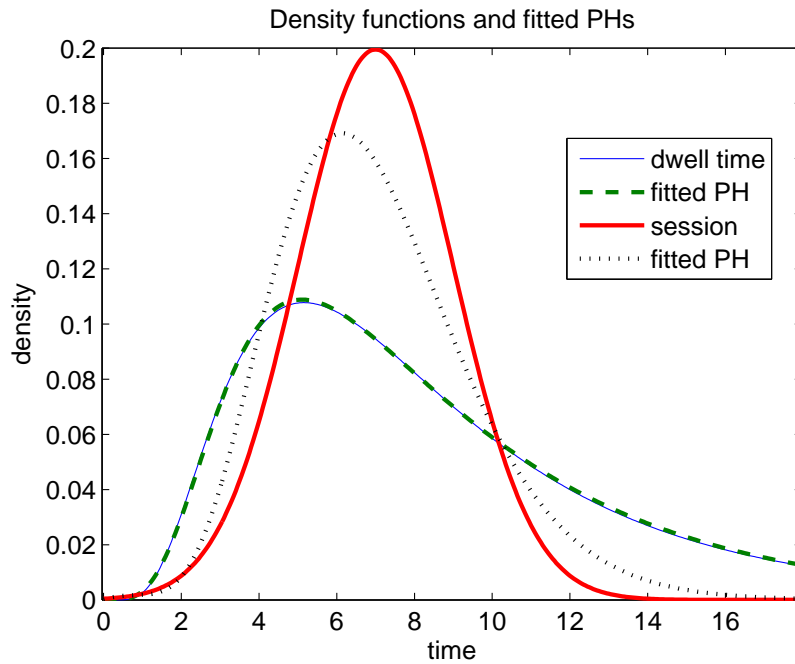


Figure 4.3: Results of PH fitting

the time consuming numerical convolutions are avoided, moreover, the resultant residual session length pdf is available in analytical form, thus useful for further analysis, while the other methods result in only series of numerical data. The Figure contains the experimental pdf of the residual session length, obtained by running a simulator of the described system as well. This curve was generated during simulating the motion and session lengths of customers in the area and after recording the remaining connection holding time of 50000 handover connections into the examined cell.

The cumulative distribution functions obtained by the four approaches are also depicted in Figure 4.5. In this graph the slight differences are less apparent, and the cdf obtained by simulation is practically equal to the one obtained by the “PH fitting dwell” approach.

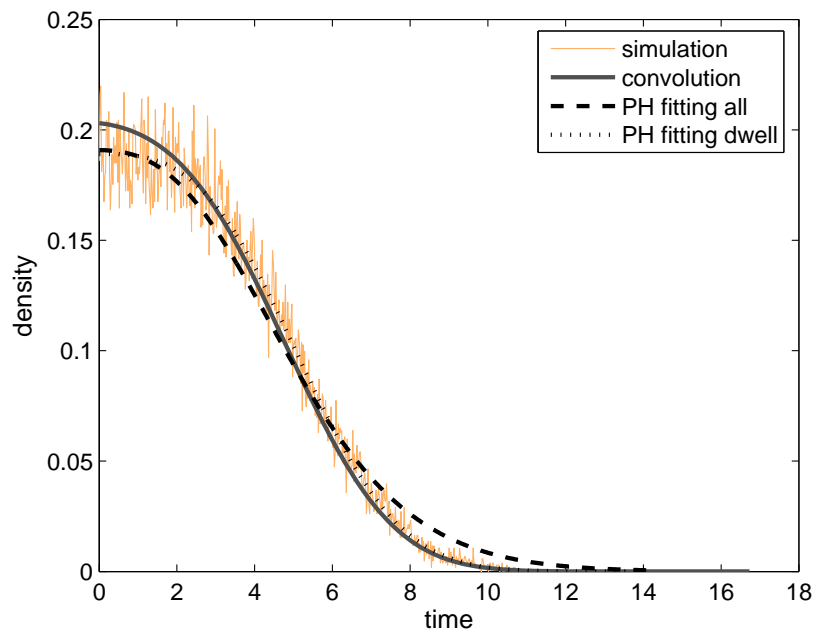


Figure 4.4: Pdf of residual session length calculated with different approaches

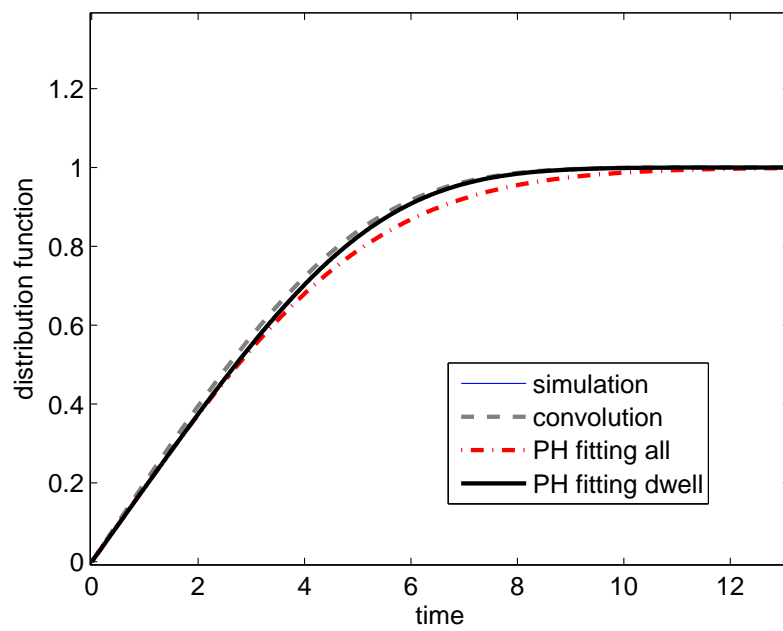


Figure 4.5: Cdf of residual session length calculated with different approaches

# Chapter 5

## Capacity and performance of 3G systems

In this Chapter I focus on the capacity issues of the 3G radio interface. The radio access technology of the UMTS (Universal Mobile Telephone System) is based on a spread-spectrum method, namely WCDMA (Wideband Code Division Multiple Access). Due to the inherent properties of CDMA access, users cause interference to each other when transmitting, or being transmitted to. This results in the phenomenon, that the capacity of the radio interface, the radio coverage and carried traffic are strongly coupled and depend on each other. The previous investigations of this Thesis were focused on the case when the capacity of the radio interface is given. This Chapter rather focuses on taking into account the effect of interference, hence examining the effective capacity of 3G systems, but the method can be directly generalised to any other CDMA system. Besides Release '99 UMTS radio interface, the HSDPA (High Speed Downlink Packet Access), defined in Release 6 of the 3G standards family is also examined.

### 5.1 3G radio interface

In Section 3.5.1 the basic operation of UMTS radio interface (in terms of usage of OVSF codes for defining physical carriers) was described. Here I investigate more on the properties of the radio interface and HSDPA extension, for the sake of better description of the issues investigated in this Chapter.

As mentioned earlier, in downlink direction the physical traffic channels of different users are separated by giving different orthogonal codes to each channel. Since these are orthogonal, theoretically the transmission to other users does not cause interference at the receiver. However, due to multipath propagation of radio signals, reflected components with different delays and

different attenuations are added to the main component at the receiver. Since OVSF codes are orthogonal only if all codes are chip-synchronous, the reflected signals lose orthogonality at the receiver, effectively causing interference. The result of this multipath effect is that each user's signal cause interference to each other user in the cell, moreover the pilot and control channels, as well as the reflected components of a given user's own useful signal cause interference. This happens even if rake receiver is applied (used to mitigate inter symbol interference). This effect is usually taken into account by means of the orthogonality factor (OF)  $\rho$ . In the literature the supposed values of  $\rho$  are between 0.5 and 0.9, accounting for bad multipath conditions (0 orthogonality would mean the complete loss of orthogonality) to good channel (OF 1 means perfect orthogonality, no multipath effect, no intra-cell interference). It is intuitively not surprising, that in practice the orthogonality factor is not constant, but rather time variant and also depends on the receiver's distance from the base station (the greater the distance, the more serious multipath effect may occur, causing the OF to decrease as function of the distance from the base station). Thus the determination of the orthogonality factor itself is addressed in the literature quite often [110][111][112][113]. In [111] the authors revealed that the median of the orthogonality factor decays with the distance according to the following simple function:

$$\rho(r) = \frac{1}{1 + \beta \cdot r}, \quad (5.1)$$

where the distance  $r$  from the base station is expressed in kilometers and  $\beta = 2.9$  value was suggested in the paper. Later in this dissertation I use this expression to describe the distance-dependency of the orthogonality factor.

It is easy to see that the OVSF spreading codes (channelisation codes) are not enough for UMTS operation. As the OVSF codes are not orthogonal if different signals do not arrive to the receiver synchronously and uplink transmission of UMTS is not synchronised among the users, OVSF codes are inherently not capable of differentiating the uplink channels of different users. Hence, another type of codes, the so-called scrambling codes are also used to separate signals of different terminals in uplink direction. These are fractions of long pseudorandom codes, having good orthogonality (cross-correlation) properties. With such a second coding of the signal (that does not spread the signal spectrum anymore, it's chiprate is the same 3.84 Mcps as that of the channelisation codes) in uplink all other terminals' signal cause interference to a particular user, with an effective interference power reduced by the processing gain, or spreading factor of the transmission applied for the user. The OVSF codes are used to create different physical channels (signalling and data channels) of a terminal. In downlink, the same quasi-orthogonal scrambling

codes are also used, but the role is to differentiate the signal of cells (sectors) that use the same carrier frequency. This means that unlike in 2G, in 3G the same carrier frequencies might be used in neighboring cells, and the cells are identified by different scrambling codes. Naturally this cause the level of interference to increase: each neighboring cell's (supposing that they use the same carrier frequency) total transmit power adds up to the total interference of a user. It is easy to see, that in both uplink and downlink, users and cells are coupled by means of the interference caused to one another: if power is increased to mitigate interference, this will force the neighbors to increase their powers. Thus, the maximum output power of the base station in downlink, and the maximum power level of users in the uplink are the limiting factors of system capacity.

In both uplink and downlink direction power control is present, to avoid unnecessarily high transmitted powers and therefore unfair access to the radio channel due to the near-far problem and to mitigate the effect of fast fading on the channel. Power control ensures that the transmitted powers are enough to maintain the necessary signal to interference ratio (SIR) for each user. Fast power control is performed in every 0.667 ms, that is in every UMTS slots. In downlink direction this means that prior to modulation each users' symbols are multiplied by a factor that results in the difference of the transmitted power targeting the given user. This means that in downlink, the total output power of the base station has to accommodate not only the power requirement of different dedicated traffic channels, but this power adjustment on each channel as well. Moreover, the output contains the power of pilot and control channels and that of the extra dedicated channels of soft-handover connections. Furthermore, if HSDPA is deployed, the shared channel carrying HSDPA traffic also requires transmission power.

As a conclusion, it is apparent that the RNC (Radio Network Controller) device of the UMTS system performs the distribution of radio resources along three dimensions. One dimension is the number of channelisation codes, allocated in discrete quantities. The other dimension of the radio capacity, the transmit time is also discrete, as the transmission time is usually allocated in 10 ms frames (or multiples, 20 ms, 30 ms). The third dimension is the transmission power, that can be viewed as a continuous resource.

### **5.1.1 HSDPA operation**

Release 6 of 3GPP standards introduced the new HSDPA (High Speed Downlink Packet Access) services. The goal was to increase the download rate of UMTS (theoretical maximum of

1920 kbps using one channelisation code, in practice typically 384 kbps) with almost an order of magnitude, to the theoretical maximum of 14.4 Mbps. The theoretical maximum term refers to the fact that the transmission rates cited here are the highest achievable bitrates of the physical layer. The motivation was that currently widespread applications generate bigger mass of downlink transmission than uplink. Latter in Release 7 the high-speed uplink extension of the standard, the HSUPA (High Speed Uplink Packet Access) was also defined.

The main features of HSDPA compared to Release '99 are the following. Transmission is accomplished on a shared channel (High-Speed Downlink Shared Channel, HS-DSCH), in contrast with the dedicated channel approach of Release '99 UMTS (although a shared transport channel is also part of the original standard). On HS-DSCH the length of channelisation codes is fixed to 16. Multiple channelization codes can be clamped together to create a single transport channel, namely a single transmission uses these codes parallelly, in order to multiply the achievable throughput. The maximum number of parallel codes is 5, 10 or 15, depending on terminal and base station capabilities. Code multiplexing might be optionally present, meaning that within a transmission interval more users are served parallelly, using different subsets of the available channelisation codes.

The channel is distributed among users in 2 ms frames (3 UMTS slots). The scheduler is placed into the base station, in order to reduce round trip time compared to original UMTS standard, where the scheduler sits at the RNC. The scheduler decides which user gets the next 2 ms time frame and chooses an appropriate transport format and send a frame on the shared channel. The transport format is described in terms of the number of parallel spreading codes used, the channel coding and the modulation. Hence, for a given transport format, the number of useful payload bits is determined. In case of HSDPA, the use of 16 QAM modulation is also allowed, resulting in the doubling of possible physical layer bitrates.

Considering the explanation of Section 3.5.1, we may conclude that a HSDPA channelisation code of length 16 provides 480 kbps physical bitrate. As HSDPA allows the concatenation of maximum 15 such codes and 16 QAM modulation, hence the 30 fold highest physical rate of 14400 kbps mentioned earlier in this Section.

Changing modulation, coding and number of channelisation codes is the channel adaptation mechanism of HSDPA. This approach is used in contrast to fast power control, to comply transmission with changing channel and interference characteristics. Channel adaptation is based on the user terminals' constant report of the channel quality. There are separate uplink control channels, where users send their experienced CQI (Channel Quality Indicator) based on their

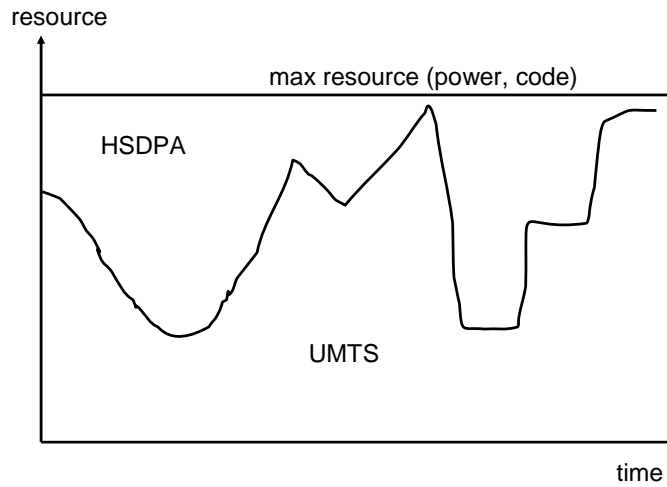


Figure 5.1: Resource sharing between UMTS and HSDPA

measurements on the common pilot channel. A given CQI value and the terminal capabilities determine the transport format the base station should send with. According to the standard, the reported CQI and corresponding transport format must enable the user to receive and decode the HSDPA frame with frame loss probability less than 0.1. The values of CQI may vary between 0 and 30, ranging from very bad channel with no chance of transmission to very high signal to interference ratio, allowing the transmission of least robust (thus highest useful bitrate) formats with maximum number of parallel codes.

In order to achieve the highest spectrum utilisation, HSDPA can operate together with conventional UMTS services on the same carrier frequency. In this case the HS-DSCH channel gets the power that is remaining after satisfying all UMTS dedicated channels and the remaining free spreading codes of length 16. Thus HSDPA operates on the remaining resources left unused by UMTS. Figure 5.1 shows this basic resource sharing policy. Analogous idea is present in 2G networks: GPRS service may use the time slots left vacant by the GSM services. As in case of 2G, where time slots might be permanently allocated for GPRS traffic only, it is possible that the operator pre-configures some resources (power, SF 16 codes) for only HSDPA use, in order to maintain a minimal HSDPA capacity. Moreover, some manufacturers' systems are not capable of dynamically allocating remaining resources to HSDPA, but fix assignment must be used, or even separate carrier frequency must be dedicated to HSDPA services. Clearly, the best spectrum utilisation is achievable in case of common HSDPA/UMTS carriers, thus my investigations will focus on this operation. However, if HSDPA traffic is present, this also increases interference level at UMTS terminals. Thus, the remaining total power cannot be allocated to HSDPA,

since the power level of UMTS traffic must also be increased due to the interference induced by HSDPA. Primarily HS-DSCH channel is transmitted with the allowed remaining power level. However, there is some form of "hidden" power control in HSDPA. Namely when the channel is very good (high CQI), the system would transmit with high bitrate transport format, it may happen that the scheduled user terminal is not capable of receiving transmission over 10 or 15 parallel codes and/or 16 QAM modulated signals. In this case the high CQI value notifies the base station to reduce its transmission power with a given amount and this reduced power is still enough for the user to decode its signal.

Along with the channel adaptation, error free transmission is achieved by means of new ARQ mechanism in HSDPA. The so-called Hybrid-ARQ uses chase combining (the erroneous packet is kept at the receiver and combined with the retransmitted one), incremental redundancy (retransmission is done with a more robust transport format) and selective repeat techniques (just damaged frames are re-sent within a transmission window) to increase robustness and efficiency of the retransmission procedure.

The standard defines 12 terminal categories. These are different in terms of their capability of receiving 16 QAM or not, the maximum number of channelisation codes they can decode, the number of soft bits they can store for turbo decoding and the required time (expressed in number of frames) between two consecutive receipt of transmit data. Figure 5.2 summarizes the most relevant features of different terminal categories. It is apparent, that the useful maximum bitrate achievable by HSDPA ranges from around 830 kbps to 12.78 Mbps, depending on the terminal category. Naturally, these rates are achievable only in case of good channel conditions and lone users. The shared nature of HS-DSCH results in the drop of individual transmission rate if other users are present. As mentioned earlier, the base station has the role of scheduling transmissions, based on channel quality measurements coming from users and past scheduling information. In the literature there are numerous HSDPA scheduling mechanisms (see e.g. [114] and references therein), but practical implementations usually consider three scheduling disciplines:

- Round Robin Scheduler: the Node B (the base station in 3G terminology) fairly schedules timeslots to users, one after the other. This scheduling is fair in terms of even distribution of timeslots among users, however unfair in terms of received throughput. Namely terminals under worse channel conditions will receive less data, due to the applicable more robust (thus lower data rate) transport formats. In case of different terminal categories, this discipline allows the better terminals (with more advanced capabilities) to gain higher throughput.

cat.	modulation	max.codes	inter-tti	eff.bitrate
1-2	QPSK/16QAM	5	3	1194.67 kbps
3-4	QPSK/16QAM	5	2	1792 kbps
5-6	QPSK/16QAM	5	1	3584 kbps
7-8	QPSK/16QAM	10	1	7205.5 kbps
9	QPSK/16QAM	12	1	8618.5 kbps
10	QPSK/16QAM	15	1	12779 kbps
11	QPSK	5	2	829.75 kbps
12	QPSK	5	1	1659.5 kbps

Figure 5.2: HSDPA terminal categories

- Proportional Fair Scheduler: this assures fair throughput distribution, as scheduling decisions are based on the past actual throughput of the users. Literally it means that users with bad channel receive more frames. If this scheduling is applied without taking into account different user capabilities, better terminals may not utilise their more advanced capabilities, as they will receive much less frames.
- Max CQI scheduler: this discipline schedules the user with the best channel conditions. This is an unfair approach, since users suffering from high interference might not be scheduled. However, this results in the highest system throughput.

## 5.2 Capacity analysis

Given the CDMA access of 3G radio interface, capacity, coverage and carried traffic are coupled and changing in time. Therefore it is not straightforward to characterize the radio interface with a single transmission capacity, given in say kbps. However, for rough capacity dimensioning purposes, both for UMTS and UMTS/HSDPA cells simple and fast methods are needed that capture the most important properties of 3G radio interface and derive results for radio capacity. The method described here takes into account user distribution, cellular topology, different transmit powers and most importantly the multiservice nature of the UMTS. What follows is a method that is applicable to answer the following questions:

- What is the average throughput of a UMTS carrier, if the available power resource is given?

- Given an average carried capacity, what is the amount of output power used? How much power is allocated to HSDPA services?
- What is the available cell throughput of HSDPA? What is the achievable user throughput over the edge of cells?
- How does these above parameters depend on user distribution, Node B powers, channel characteristics, interference level, etc.?

### 5.2.1 System description

The forthcoming analysis is carried out supposing the system described in this Section. Basically the analysis will concentrate on a given cell, however the effect of surrounding cells will be considered. I suppose omnidirectional transmit antennas and terrain without mayor obstacles, resulting in circular cell areas, with the base station in the middle of the circle. Due to the interference limited nature and coupled coverage and traffic of CDMA networks, usually cell size in UMTS is much smaller than in GSM. The radius of the examined cells is denoted with  $R_{cell}$ . The placement of neighboring base stations are determined according to standard hexagonal layout, the influence of two rings of neighboring cells (18 base stations) will be considered.

The interference effect of neighbors are calculated based on channel attenuation model. Channel is considered to influence the signal power with exponential path loss, generally

$$PL(r) = \beta \cdot r^{-\gamma}. \quad (5.2)$$

This exponential model incorporates well known channel models, such as COST 231, Okumura-Hata or Hata models, where different propagation environments (e.g. rural, suburban, urban), carrier frequencies and topological parameters (base station height, mobile antenna height) are considered via the parameters  $\beta$  and  $\gamma$  in (5.2). Neighboring base stations are modelled as source of interference power, transmitted at constant level, but unlike most of the literature, here base station powers may be different. Thermal noise will be basically neglected in numerical calculations, but shown in basic theoretical expressions. The reason of this is that at room temperature, the noise has power of around  $-107$  dBm at the 5 MHz wide UMTS band. Considering the cellular layout described above, 1 W of the furthest interfering base station would cause around  $-117$  dBm interference power at a customer that is placed into the far end of the considered radio cell. The noise power hence is just 10 dB above this itself negligible interference. More-

over, we are rather interested in how interference influences system behaviour. Nevertheless, the incorporation of thermal noise into the calculations is straightforward.

I assume that UMTS users and HSDPA users are present in the cell and a single carrier accommodates both traffic. Users of UMTS might use any of  $K$  service classes (bearers). Based on the chiprate and modulation described earlier and on the possible lengths of OVVSF codes, the physical transmission rate in downlink direction may vary between 15, 30, 60 ... 1920 kbps, corresponding to the use of channelisation codes of length 512, 256, 128 ... 4 chips. However, in practical systems just a few services (radio bearers) are used. The typical services have useful data rate (not including the redundancy induced by forward error correction and necessary signalling bits) of 12.2 kbs (maximum rate of the variable bitrate AMR speech codec), 64 kbps (low resolution video, high quality audio or low rate data), 144 and 384 kbps bearers (video or data). The corresponding spreading factors are 128, 32, 16 and 8 (the amount of redundancy is notable, considering the corresponding physical data rates of 60, 240, 480 and 960 kbps).

Each service class is characterised by the signal to interference ratio required for acceptable performance. This is more precisely expressed in terms of required bit energy per interference spectral density ratio  $\frac{E_b}{I_0}$ , for now on the  $\frac{E_b}{I_0}$  requirement is denoted by  $\epsilon_k$  for service type  $k$  or user  $k$ . The values of  $\epsilon_k$  are assumed to be known. In the literature, several sets of values can be found, based on link-level simulations or just estimations. However, as it will be shown, the performance of the system is highly dependent on these values, thus in a real capacity planning procedure should be chosen very carefully.

User distribution is assumed to be known over the cell area and given in terms of the pdf of user positions. However, it is reasonable to assume that users of different services have different distribution over the cell disc (e.g. HSDPA users are rather found on places that represent residential area, while UMTS voice bearer users rather dwell on street area, etc.). Thus different distribution for different service class users may be assumed. Usually the basic and most common assumption is the even distribution of users over the area, but this assumption will be released in this dissertation.

### **5.2.2 Capacity of Release '99 radio interface**

In order to analyse the 3G operation with HSDPA deployed, first a cell with only conventional UMTS traffic is analysed. The following investigation is based on the basic downlink SIR equation of UMTS systems (e.g. [19], but naturally all papers use some form of these basic

equations). To investigate the capabilities of the 3G radio interface, the basic signal to interference ratio equation should be formulated, for a given user  $i$ , served by base station 0. Namely, approximating interference as Gaussian, the bit energy over interference spectral density ratio should be over a given threshold, to achieve the necessary bit error ratio for a service. Given the user is placed on polar coordinates  $(r_i, \phi_i)$  with the serving base station in the origo this is:

$$\left[ \frac{E_b}{I_0} \right]_i = \frac{Rc}{Rb_i} \cdot \frac{P_i^0(r_i, \phi_i) \cdot L^0(r_i)}{(1 - \rho(r_i)) \cdot P_{\text{inst}}^0 \cdot L^0(r_i) + \sum_{b \neq 0} P_{\text{inst}}^b \cdot L^b(r_i, \phi_i) + P_{\text{noise}}} \geq \varepsilon_i, \quad (5.3)$$

where  $P_i^0(r_i, \phi_i)$  is the transmission power targeted to user  $i$ ,  $L^0(r_i)$  is the channel gain between the serving base station and the user, depending on the user's distance  $r_i$ ,  $P_{\text{inst}}^0$ -s is the total instantaneous transmitted power of the serving base station,  $P_{\text{inst}}^b$  denotes the instantaneous power radiated by the  $b$ th neighboring base station,  $L^b(r_i, \phi_i)$  is the path gain between this and user  $i$ . The power of the pilot and control channels is included in  $P_{\text{inst}}^0$ , thus its power is

$$P_{\text{Pil}}^0 = P_{\text{inst}}^0 - \sum_i P_i^0. \quad (5.4)$$

$\rho(r_i)$  is the orthogonality factor,  $P_{\text{noise}}$  is the power of the thermal noise.  $Rc$  and  $Rb_i$  are the chiprate and bitrate of the given service used by the customer. Regarding the latter quantity, in some cases it is the useful bitrate of the service, that is much lower than the actual physical bitrate. In other interpretation  $Rb_i$  is the physical symbol rate, in this case the fraction  $Rc/Rb_i$  (processing gain) is equivalent with the spreading factor (length) of the channelisation code used for the given service. Note that any interpretation might be used analogously, the difference results in the change of the required  $\varepsilon_i$  threshold. In the following I use the former representation.

It is worth noticing here, that inequality (5.3) does not contain antenna gain, however if supposed, this should be simply incorporated in the expression of the path gain as a multiplier factor (or an additive factor on dB scale). Moreover, although in this analysis omnidirectional antennas of base stations were supposed, the incorporation of sectorized antennas is easy at this step. Namely a horizontal antenna characteristic should be used (that gives the extra attenuation of the antenna, as the function of the angle between the main direction and the position we are interested in). This again can be incorporated into the expression of path gain, but now this will depend on not only the distance from the base station, but on the direction of the position vector, i.e.  $L^0(r_i)$  is replaced by  $L^0(r_i, \phi_i)$ . The direction-dependent extra attenuation should also be taken into account when calculation interference power from neighboring base station.

Inequality 5.3 must hold for all active connections. By examining relationship (5.3) one can see how the capacity and traffic is coupled: any additional transmission will raise the interference

power at the denominator, requiring the raise of the useful power at the nominator, that results in the raise of transmitted powers to all users. The amount of traffic is bounded by the finite transmission power of the base station, namely

$$\sum_i P_i^0(r_i, \phi_i) + P_{\text{Pil}}^0 \leq P_0^0. \quad (5.5)$$

The code dimension of downlink radio resource also bounds the number of parallel connections. Referring to Section 3.5.1, the code constraint can be formulated as

$$\sum_k n_k \cdot 2^{9-k} \leq 504, \quad (5.6)$$

where  $n_i$  is the instantaneous number of connections using channelisation code of length  $2^i$ . This expression is valid for data traffic channels. As the primary and secondary Common Control Physical Channels (CCPCH) and the Common Pilot Channels (CPICH) typically occupy 4 OVSF codes of length 256, data physical channels might occupy 504 at the bottom level of the code tree, hence this number in (5.6) instead of 512.

The problem of the multi-service nature of UMTS is taken into account by means of different  $\varepsilon$  requirements and the different processing gains of services, hence the power required for a service depends not only on its distance from the base station.

The problem of determining the capacity of such a multiservice network is arisen because of the high number of possible combinations of active users of different service types, not to mention the problem of their position. Therefore I propose the following method.

I define the average capacity of a UMTS carrier as following. Let the useful bitrate of the  $k$ th bearer type be denoted by  $Rb_k$  and the average number of active bearer  $k$  is  $N_k$ . This latter average is understood as, supposing any given user traffic profile and service mix, the average number of type  $k$  bearers scheduled in every radio frame is  $N_k$ . Clearly, this quantity depends on the traffic demand as well as scheduling policy of the network and user positions in the network (as this basically influences required power). Given these circumstances the average capacity of a UMTS carrier is defined as

$$\bar{R}_{\text{UMTS}} = \sum_{k=1}^K N_k \cdot Rb_k. \quad (5.7)$$

In the following, we will use the average number of total scheduled connections,  $N$  and the ratio of type  $k$  bearers  $n_k$ , clearly  $N_k = N \cdot n_k$  and

$$\bar{R}_{\text{UMTS}} = N \cdot \sum_{k=1}^K n_k \cdot Rb_k. \quad (5.8)$$

This quantity characterizes the average amount of useful throughput carried over a UMTS carrier, therefore the term capacity is used. However, because the coupled nature of traffic and capacity, this may not be straightly seen as a given capacity to be shared (but from a system level point of view, a given cell can be seen offering this capacity on average) among users. In the following, the term capacity and throughput will be used interchangeably for this quantity.

This quantity is very useful during network dimensioning phase. At this stage the operator needs rough estimate of the average traffic that can be transmitted over an area with a given number of cells (in order to estimate the required number of cells). As we will see, that this defined average capacity takes into account the multiservice nature of UMTS, as well as radio conditions (path loss) and interference.

To obtain carrier throughput, the following calculations should be performed. After rearranging (5.3) and considering that power control forces the system to achieve, but not to exceed the SIR requirement (i.e. the equality is supposed), we get:

$$P_i^0(r_i, \phi_i) = \varepsilon_i \cdot \frac{Rb_i}{Rc} \cdot \left( (1 - \rho(r_i)) \cdot P_{\text{inst}}^0 + \sum_{b \neq 0} P_{\text{inst}}^b \cdot \frac{L^b(r_i, \phi_i)}{L^0(r_i)} + \eta_i \right), \quad (5.9)$$

where  $\eta_i = \frac{P_{\text{noise}}}{L^0(r_i)}$  is the relative noise power. Using the expression of  $P_{\text{inst}}^0$  from (5.5) this expression results in the linear system:

$$P_i^0(r_i, \phi_i) = \varepsilon_i \cdot \frac{Rb_i}{Rc} \cdot \left( (1 - \rho(r_i)) \cdot \left( P_{\text{Pil}}^0 + \sum_j P_j^0(r_j, \phi_j) \right) + \sum_{b \neq 0} P_{\text{inst}}^b \cdot \frac{L^b(r_i, \phi_i)}{L^0(r_i)} + \eta_i \right), \quad (5.10)$$

The latter sum in the right hand side of (5.10) draws special attention in the literature. Namely, considering all neighboring base station transmits with the same power as the examined base station, the sum

$$\sum_{b \neq 0} \frac{L^b(r_i, \phi_i)}{L^0(r_i)} = f(r_i, \phi_i) \quad (5.11)$$

is called other to own cell interference factor or geometry factor. In our analysis it is not required that all neighboring Node B-s transmit with the same power, although it is convenient to introduce the notion of other to own cell path gain factor, however in this case this describes the effect of only one interfering Node B. That is:

$$f^b(r_i, \phi_i) = \frac{L^b(r_i, \phi_i)}{L^0(r_i)}. \quad (5.12)$$

Let us concentrate on the mean of transmitted powers in the system. If we take the expectation of (5.9), this result in the average power transmitted to a user. This average will be the same for

all terminals using a given bearer type  $k$  (as the parameters affecting the average are the bearer type dependent processing gain, SIR requirement and spatial distribution). Consequently, the following expression is viewed as the average transmitted power to a type  $k$  connection. That is

$$\overline{P}_k^0 = \varepsilon_k \cdot \frac{Rb_k}{Rc} \cdot \left( (1 - \overline{\rho}_k) \cdot P_{\text{avg}}^0 + \sum_{b \neq 0} P_{\text{avg}}^b \cdot \overline{f}_k^b + \overline{\eta}_k \right). \quad (5.13)$$

The average orthogonality factor depends on the service class, since user distributions of each service class  $k$  might be different, the same argument applies to the other to own cell path gain factor. In the expression above,  $P_{\text{avg}}^0$  and  $P_{\text{avg}}^b$  denote the average output powers of the examined base station and that of the interfering base stations. The average orthogonality factor can be calculated as

$$\overline{\rho}_k = \int_0^R \int_0^{2\pi} \rho(r) \cdot g_k(r, \phi) d\phi dr, \quad (5.14)$$

where  $g_k(r, \phi)$  is the probability density function of class  $k$  user distribution over the cell, expressed on a polar coordinate basis. It is obvious, that if the user distribution is given in cartesian coordinates (and the corresponding density function as well), first the coordinate transform should be performed to get  $g_k(r, \phi)$ . As an example, if users are evenly distributed over the disc representing a cell, the density function on cartesian coordinates is  $g_k(x, y) = \frac{1}{R^2\pi}$ , after transforming the density function in polar coordinates is  $g_k(r, \phi) = \frac{r}{R^2\pi}$ . Sticking to the example of evenly distributed users and using the distance dependency of the orthogonality factor (5.1), for the average we get:

$$\overline{\rho}_k = \frac{2}{\beta^2 \cdot R^2} (\beta R - \ln(1 + \beta R)), \quad (5.15)$$

where  $R$  is the cell radius. One can observe, that – in contrast with the common assumption of having constant orthogonality factor – the cell radius will have impact on performance, through the average OF.

The average relative noise power is similarly calculated:

$$\overline{\eta}_k = \int_0^R \int_0^{2\pi} \frac{P_{\text{noise}}}{L(r)} g_k(r, \phi) d\phi dr, \quad (5.16)$$

if we assume even user distribution and exponential path loss model with parameters  $\beta$  and  $\gamma$ , this results in

$$\overline{\eta}_k = \frac{2 \cdot P_{\text{noise}}}{\beta \cdot \gamma + 2\beta} R^\gamma. \quad (5.17)$$

With regards to the other to own cell path loss ratio, the following can be stated. According to Figure 5.3, with a customer dwelling at point  $(r, \phi)$ , supposing exponential pathloss model with  $\gamma$  exponent:

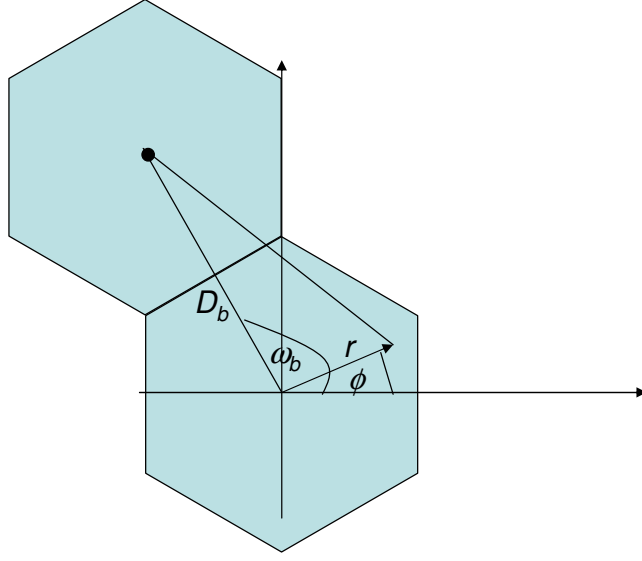


Figure 5.3: Distances for calculating  $f_k^b(r, \phi)$

$$f_b^k(r, \phi) = \left( \frac{D_b^2}{r^2} - \frac{2D_b}{r} \cos(\omega_b - \phi) \right)^{-\frac{\gamma}{2}}, \quad (5.18)$$

where  $D_b$  is the distance between base station  $b$  and  $0$ ,  $\omega_b$  is the angle between the line connecting these two base stations and the  $x$  axis. Based on this, the average other to own cell path loss ratio is calculated as:

$$\overline{f_k^b} = \int_0^R \int_0^{2\pi} f_k^b(r, \phi) \cdot g_k(r, \phi) d\phi dr. \quad (5.19)$$

After determining the required mean parameters, equation (5.13) must be solved for all  $K$  service classes with a supposed level of average used power ( $P_{\text{avg}}^0$ ). As result, we have average power levels of all  $K$  bearer types, namely  $\overline{P}_k^0$ .

The average number of simultaneous transmissions at the radio interface was earlier introduced and denoted by  $N$ . As stated earlier, this means that on a long time average, there is  $N \cdot n_k$  active connections of type  $k$  on the radio interface. This inherently means, that the average used output power is

$$N \cdot \sum_{k=1}^K n_k \cdot \overline{P}_k^0 + P_{\text{Pil}}^0 = P_{\text{avg}}^0. \quad (5.20)$$

After the average power levels are calculated from (5.13),  $N$  has to be determined from (5.20). After having  $N$ , the average useful capacity (or average useful throughput using a given average power level) of the UMTS bearer is simply calculated as (5.8).

For the further capacity evaluation of 3G radio interface, with HSDPA enabled, we need another measure of the system. As it was outlined earlier, HSDPA may use the remaining transmission power of the base station, thus it is required to somehow characterize the used power of Release '99 transmission. The specific question to be answered is "what is the average used power for Release '99 transmission, given an average Release '99 traffic amount of  $R_{UMTS}$  kbps". The former idea can be used again in the reverse direction. Namely, for obtaining the used power,  $N$  is expressed from (5.8), with the given  $R_{UMTS}$  traffic. Then, since in this case  $P_{avg}^0$  is unknown, it's expression from (5.20) with the previously determined  $N$  should be substituted into (5.13). Then we arrive to a slightly modified version of the basic equations, namely

$$\overline{P}_k^0 = \varepsilon_k \cdot \frac{Rb_k}{RC} \cdot \left( (1 - \overline{\rho}_k) \cdot (N \cdot \sum_{l=1}^K n_l \overline{P}_l^0 + P_{Pil}^0) + \sum_{b \neq 0} P_{avg}^b \cdot \overline{f}_k^b + \overline{\eta}_k \right). \quad (5.21)$$

This linear system has to be solved for all  $\overline{P}_k^0$ -s, then the resultant average used power is given as in (5.20).

When performing this calculation, some notes on the power of interfering base stations should be given. One method of modeling the neighboring interference power is simply to use a constant value, given in Watts and substitute this into the calculations. By choosing this value to be the largest possible power, a worst case scenario can be evaluated. The former investigations allow the use of different powers in neighboring base stations as well.

However, an important and realistic scenario is when the neighboring base stations carry about the same amount of traffic as the examined one. If we want to calculate the average used power for accommodating a given amount of traffic, this raises the problem that interfering Node B power should appear in equations (5.21) that is only available after solving it. This implies the use of the following iterative approach.

- Step 0. Suppose arbitrary level of interfering Node B powers (less than the maximal output power).
- Step 1. With the given interfering power level solve (5.21). Determine the used power of the Node B in question with (5.20).
- Step 2. Substitute the resultant power level as interfering power. Repeat step 1 and step 2 until convergence.

Convergence is achieved when the powers calculated in two successive iterations do not differ. It is straightforward to extend this iterative approach, when the neighboring base stations carry

unequal traffic. In this case the calculation of used power should be performed in all cells, one by one, with using interfering powers calculated earlier for neighboring cells. The iteration goes "round" all the cells until convergence is achieved.

### 5.2.3 Capacity and performance of HSDPA

To evaluate the average capacity of a 3G carrier when HSDPA is deployed, the basic SIR equation is the starting point, written for the HSDPA channel. For a HSDPA user  $i$ , served by base station 0 that is

$$SIR_{HS}(r, \phi) = 16 \cdot \frac{P_{HS}^0 \cdot L_i^0(r, \phi)}{(1 - \rho(r)) \cdot P_{inst}^0 \cdot L_i^0(r, \phi) + \sum_{b \neq 0} P_{inst}^b \cdot L_i^b(r, \phi) + P_{noise}}. \quad (5.22)$$

In this expression the term 16 is the fixed spreading factor used in the HSDPA channel. Given that HSDPA uses multicode transmission, each physical channel (spreading code) causes interference to the other ones due to multipath propagation, hence the power of HSDPA also appear in the denominator of (5.22) as the part of the total output power. Considering this, the only change regarding the former equations is that the HS-DSCH channel appears as if it was a new type of UMTS service class, but it *does not have an explicit SIR requirement*, rather it chooses transport format according to the SIR of the channel.

As it was outlined earlier, the HSDPA terminal reports it's perceived SIR to the base station by means of the CQI parameter, that – along with the terminal category – explicitly determines the transport format it can receive. Thus, in order to evaluate HSDPA in terms of transmission capacity, a relationship is needed which connects SIR with CQI. In general this depends on the quality of the receiver at the terminal – if a manufacturer is able to produce a receiver that can decode the least robust transport format at very low SIR, the terminal should report high CQI. Therefore it is very hard to find general SIR-CQI relationship. In fact, most of the literature rely on [115], where, based on detailed link-level simulations, the following relationship was found for 0.1 block error probability:

$$CQI = \left\{ \begin{array}{l} 0 \\ \lfloor \frac{SIR}{1.02} + 16.62 \rfloor \\ 30 \end{array} \right\} \begin{array}{l} SIR \leq -16 \\ -16 < SIR \leq 14 \\ SIR > 14 \end{array} \quad (5.23)$$

where the SIR should be given in decibels. Another expression found for this relationship is given in [116], that is simply

$$CQI = \lfloor SIR + 3.5 \rfloor. \quad (5.24)$$

By observing the two expressions we see that the term corresponding to the spreading factor 16 is the difference, namely the definition of SIR differs in this term in the two expressions, otherwise they are identical.

The basic method to evaluate HSDPA is to assign a single power value to the interfering base stations and another value for the Release '99 ( $P_{\text{UMTS}}^0$ ) traffic within the cell, and set HSDPA power ( $P_{\text{HS}}^0$ ) to a constant level (this models HSDPA systems with pre-configured HSDPA power). Then in (5.22) the term  $P_{\text{inst}}^0 = P_{\text{UMTS}}^0 + P_{\text{HS}}^0 + P_{\text{Pil}}^0$ . Obtaining the SIR from this, transforming SIR into CQI with (5.24) and using the CQI-transport format tables given for each terminal categories in the standard, the following simple questions can be answered

- what is the available transmission rate of a terminal of given category placed at a specific location
- what is the available throughput of a given terminal at the cell border
- what is the average cell throughput achievable by HSDPA.

For obtaining the average capacity, the following method is proposed.

As it is outlined earlier, the instantaneous SIR a terminal perceives determines the CQI it is reporting to the base station, with the relation defined by (5.23) or (5.24). The standard contains tables that define the applicable transport format when a given CQI is reported, for each terminal category. The transport format explicitly determines the transmitted useful bits during a frame, hence the instantaneous useful transmission rate. Thus, from (5.22) the achievable throughput of a terminal of category  $i$  can be directly derived, using the mapping

$$\text{SIR}_i(r, \phi) \Rightarrow \text{CQI}_i(r, \phi) \Rightarrow R_i(r, \phi), \quad (5.25)$$

where  $R_i(r, \phi)$  is the achievable transmission rate of a category  $i$  terminal at position  $(r, \phi)$ .

Earlier we supposed that user distribution is given in some form over the disc representing the cell area, let us now denote the density function of HSDPA users' position by  $g_{\text{HS}}(r, \phi)$ . The average throughput achievable by the category is then calculated as:

$$R_i = \int_{r=0}^R \int_{\phi=0}^{2\pi} R_i(r, \phi) g_{\text{HS}}(r, \phi) d\phi dr. \quad (5.26)$$

We assume that there is information on the ratio of different terminal categories among all the HSDPA devices, that is  $\alpha_i, i \in [1..12]$  for category  $i$ . The average cell capacity of HSDPA is

then given by averaging over the device categories:

$$R_{\text{HSDPA}} = \sum_{i=1}^{12} R_i \cdot \alpha_i. \quad (5.27)$$

This quantity is well characterizing the HSDPA service in terms of the average amount of traffic it can serve, if the operator configures a given maximum level of radio resource (codes and power) to HSDPA.

To take deeper insight into system performance we have to handle the case when HSDPA service uses the power and code resource that is left unused by the Release'99 traffic. That is, the HSDPA resource is not supposed to be known in advance, but it is calculated from the Release'99 traffic amount.

In this case the following method is applicable. The aim is to determine the average available power and with this, the achievable average throughput of HSDPA services, given that the cell is carrying an average of  $R_{\text{UMTS}}$  kbps UMTS traffic. The used power of UMTS has to be determined basically according to the method described above. The difference is, that HSDPA traffic also means intra-cell interference for Release'99 traffic and vice versa, thus the linear system of equations (5.21) is changed. As we are interested in HSDPA capacity and HSDPA utilizes the total remaining power of the Node B, the modified expression contains all the maximum possible output power of the Node B, namely

$$\bar{P}_k^0 = \varepsilon_k \cdot \frac{Rb_k}{Rc} \cdot \left( (1 - \bar{\rho}_k) \cdot (P_0^0) + \sum_{b \neq 0} P_{\text{avg}}^b \cdot \bar{f}_k^b + \bar{\eta}_k \right), \quad (5.28)$$

and the average HSDPA power is coming from the last equation of the system, that is the one that assures that all achievable power is allocated, namely

$$\bar{P}_{\text{HS}}^0 = P_0^0 - N \cdot \sum_{k=1}^K n_k \cdot \bar{P}_k^0 - P_{\text{Pil}}^0. \quad (5.29)$$

As earlier, during this task the average number of "virtual" Release'99 connections is given through the average useful traffic  $R_{\text{UMTS}}$ , namely

$$N = \frac{R_{\text{UMTS}}}{\sum_{k=1}^K n_k \cdot Rb_k}. \quad (5.30)$$

It is easy to see from the (5.21) and (5.28) what was explained earlier: including HSDPA traffic will increase the necessary power level for UMTS, thus allowed HSDPA power is not just the difference between the total output power and UMTS power.

Obtaining average HSDPA throughput and mean achievable HSDPA user throughput on a location again simply means determining the SIR from (5.22), with the average  $\overline{P}_{\text{HS}}^0$  used and then applying (5.24) and the averaging technique described above.

### 5.3 Numerical evaluation

In this Section the average capacity of 3G systems is evaluated, based on the numerical methods investigated in previous sections. The analytical examinations are compared with results of snapshot simulations. The latter is based on uncorrelated random snapshots of the possible traffic scenarios.

During a snapshot, users placed within the cell randomly, according to user distribution over the surface. For each customer the service class  $k$  is chosen with given probability  $n_k$ . The system (5.10) is solved for a current set of users, and the total used power as well as the number of occupied spreading codes is determined. The placement of new customers is continued one by one and with each new customer (5.10) is solved again, until the total used output power reaches a given level or the number of occupied codes gets greater than the maximum. The placement is stopped at this moment.

Here it is important to note that the aim of the simulations is not to investigate some advanced admission control, load balancing or scheduling, etc. mechanisms, but to provide results in terms of average used power and average cell throughput, or user distribution if necessary, in order to compare these with the results coming from analytical investigations. Therefore just two very simple simulation policies are applied during the investigations:

- In the conservative processing method, the last customer, with whom the power level or number of spreading codes exceeded the maximum is ignored. Snapshot statistics (used power levels of customers, actual position of customers, total used power, total used spreading codes, total useful transmission rate of customers, ratios among service classes in the current snapshot) are collected for the previously admitted users, discarding the last one.
- The second processing method is to include the last user into the statistics as well. As user generation stops when the given amount of maximum power is exceeded, this method is unrealistic since it may let the used power to exceed the maximum. However, if this level

is set to be less than the physical maximum power of the base station, most snapshots will result in feasible total transmission power level.

With the first processing policy the simulation is realistic in terms of not letting more than the given maximum amount of transmission power and spreading codes to be used. However, it is apparent that this policy will bias the actual statistics compared to the parameters given as inputs to the simulations. Namely users' actual spatial distribution will change, as it is likely that a distant user is discarded due to its required higher power level (hence the actual distribution will be sparser in the vicinity of cell border, compared to the given input distribution). Same argument applies to the actual service class probabilities: higher bitrate (thus lower processing gain) connections will be often the last ones to be discarded due to their required higher power levels. This result in the distortion of actual service class ratios. The average used power level is naturally less than the given maximum level.

By applying the second simulation method, neither the spatial distribution nor the service class probabilities of customers will be distorted, as all generated users will be served. However, if the last customer is also accounted, the actual total power level might far exceed the given maximum (and can be easily greater than the typical maximum of 30 Watts of Node Bs available on the market). In most scenarios, where users are spread around the base station, usually the power is the bottleneck resource, namely user generation during a snapshot stops because of lack of base station power. However, in scenarios where customers are likely to appear near the base station, thus experiencing less interference, the code resource may become bottleneck, hence the OVSF code tree "runs out" before the power. This would mean that the customer generation is stopped because the code limit is exceeded. Considering the second processing policy, the actual average power level or even the average used spreading code number might be greater than the given thresholds. To overcome these shortages, typically smaller threshold values should be given, than the actual physical thresholds. Another problem with this method is the possibility of reaching the pole of power equations. Namely the possible solutions of the linear system (5.10) might get values that result in extremely high total power, or even negative powers. To see this we should sum up (5.10) for all users  $i$ , hence the left side of the resulting sum is the total used power, without pilot and signalling channels:

$$P_{\text{inst}}^0 = \frac{\sum_i \varepsilon_i \frac{Rb_i}{Rc} \left( (1 - \rho(r_i)) P_{\text{Pil}}^0 + \sum_b P_{\text{inst}}^b f^b(r_i, \phi_i) \right)}{1 - \sum_i \varepsilon_i (1 - \rho(r_i)) \frac{Rb_i}{Rc}}. \quad (5.31)$$

The nominator of this expression can be viewed as the downlink loading factor, as this expression approaches 0, the total used power approaches infinity. Moreover, the downlink loading

factor can easily become negative (in case of small OFs and small processing gains, hence with high-speed connections and distant mobiles in case of large cells). Therefore, if during the simulation, the last customer is also taken into account in the statistics (with whom the total output power exceeded the given maximum level), this may cause that an extreme high level of power is accounted. Nevertheless, this method is also applicable to test the accuracy of the numerical method.

Taking the above reasoning into account the conservative processing method is used during the simulations, despite its biasing effects. As the analytical method presented in previous Sections require the knowledge of user spatial distribution, service class probabilities and average used powers, to evaluate the validity of this method in an actual scenario, first the snapshot simulation should be run to obtain actual parameters. The actual user spatial distribution is collected numerically, in terms of experimental probability density function of the actual distances from the base station, per service class, and used in the analytical calculations. In this manner the biasing effect of the snapshot simulation is overcome.

It is very important to point out that during the investigations the role of the simulations is to obtain the actual parameters needed for the calculations and to justify the analytical method. Apparently no specific assumptions on actual detailed data traffic characteristics, scheduling principles, etc. are used, nor do we simulate system behaviour in details. For this purpose the two basic snapshot simulation methods are sufficient. From the presented point of view, specific radio resource management operation, based on traffic characteristics would result in the change of average used power level and the service mix, thus the propose method will be basically applicable for evaluating other scenarios.

In real life situations, the proposed methods may be used for rough dimensioning purposes. This phase of radio network planning precede the detailed cell planning and its main output is the number and position of cells to cover a given area with. This means that inherently the parameters available for this task are also rough estimations. User spatial distribution might be estimated according to population statistics and anticipated penetration of UMTS usage of users, or based on previous experiences on 2G networks. Service usage ratios may be estimated based on anticipated popularity of different services among users (this can be significantly influenced by pricing) and on the knowledge of the basic scheduling and load distribution mechanisms implemented in the actual network devices, or experiences of already operating network segments. Development of the method presented in this Chapter was motivated by actual network dimensioning problems and was conducted in collaboration with a hungarian 3G operator. The method

was partially implemented as part of a network dimensioning software tool, that is currently used by the operator.

### 5.3.1 Capacity of UMTS system without HSDPA service

The first set of investigations target a UMTS cell without HSDPA services. The effect of used power, user spatial distribution, service class ratios, cell size and distance-dependent orthogonality factor are revealed. In all the calculations 4 service classes are assumed, with 12.2, 64, 144 and 384 kbps useful throughputs. The requested signal to interference ratios for the services are assumed to be 8, 7, 3.5 and 4 decibels, respectively. The channel path loss is taken into account with exponent 4.8 (this model approximates the Okumura-Hata path loss model fairly accurately). In the basic setting two rings of neighboring base stations (18 Node Bs) were accounted as interferers, with equal transmission powers of 21 Watts. The power necessary for pilot and signalling channels was set to constant 3 Watts.

In general, three basic service mixes are investigated, namely the ratio of 12.2, 64, 144 and 384 kbps connections are assumed to be 0.4 0.3 0.2 and 0.1 in service mix 1; 0.3 0.3 0.2 and 0.2 in service mix 2; 0.2 0.3 0.2 and 0.2 in service mix 3, that is we examine the effect of increasing the amount of highest bit-rate connections at speech connections expense.

Figure 5.4 displays the average achievable cell throughput as cell radius is increased, supposing even user distribution over the cell. The two set of results compare the case of constant orthogonality factor ( $\rho = 0.7$ ) and distance-dependent OF according to (5.1). The graphs were obtained with simulations according to the second processing policy, with threshold power usage of 20 Watts. It is apparent, that with constant orthogonality factors the cell size does not influence the achievable throughput. This is because during this calculations the thermal noise was neglected. The effect of white noise would become significant in case of large distances from the Node B, hence with noise, the average cell capacity would decrease as cell size increases. The same applies to the used average power levels, as revealed by the Figure (power levels are obtained by means of simulation and were input to capacity calculations). Though the decrease of average OF forces higher level of average used power, this is still not enough to stop the decrement of cell throughput in the variable OF case. It is apparent, that calculating with the distant variable nature of the OF reflects the experience that smaller cells have higher capacity, in contrast with the constant OF resulting in cell size independent capacity.

Figure 5.4 indicates that in case of applying the distance-dependent orthogonality factor

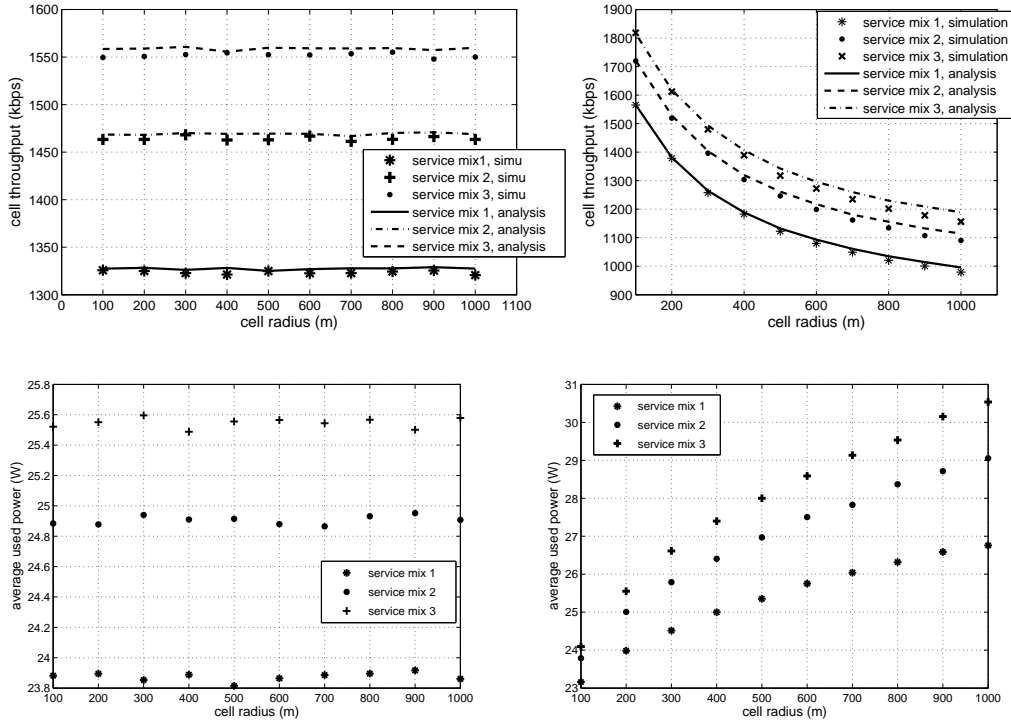


Figure 5.4: Average cell throughput (first row) and used power (second row), with constant (left) and distance-dependent (right) orthogonality factors

model, the accuracy of the numerical method decreases when considering bigger cell sizes, especially for service mix 3. Figure 5.5 plots the accuracy measure for the distance-dependent orthogonality factor scenario. Accuracy is defined as the ratio of the difference of results (analysis and simulation) and the results obtained by simulation. This increasing inaccuracy is due to the fact that was explained regarding the snapshot simulation: small orthogonality factors and the relatively higher ratio of 384 kbps services allow the cell to approach its pole, and in particular snapshots the actual power level gets very high, causing the average used power to increase. The analytical model reflects this increment in terms of a bit higher number of users, hence total cell throughput. The case of constant orthogonality factor does not have this effect, hence the accuracy of the numerical method remains under 0.005 for all examined cell sizes in that case (not plotted).

The proposed capacity evaluating method is also useful in determining the effect of user distribution in the cell. To capture this effect, besides the even user distribution over the cell (referred to in the following as "even scenario") two others were considered:

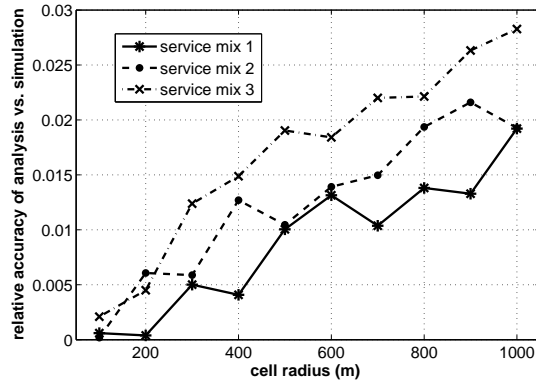


Figure 5.5: Accuracy of the analysis in case of even user distribution, distance-dependent orthogonality factor

- "hotspot scenario": user distribution over the plane follows a symmetric two dimensional normal distribution (with independent normally distributed  $x$  and  $y$  coordinates, with means 0 and equal variances), this is used to model the case when users are mainly placed around the base station
- "concentrated scenario": asymmetric two dimensional normal ( $x$  and  $y$  coordinates are independent normally distributed variables, with different means and variances) user distribution that models the case when users are mainly localized in a well defined small area, apart from the base station.

Naturally both distributions are truncated and normalized to be based over the cell.

Using these latter two scenarios (and especially the hotspot scenario) and the second simulation policy would often lead to reaching the pole of the system, due to the high number of admitted users near the base station (the proximity of the base station mean low inter-cell interference and high orthogonality). Therefore the following investigations were carried out using the first simulation policy.

Figure 5.6 compares the effect of user distribution: it is apparent that in the hotspot case cell capacity is more than the double of the capacity in the evenly distributed scenario. Bigger capacity of the hotspot scenario is not a surprise, as users generally enjoy higher SIR values as they generally dwell around the base station.

The accuracy of the proposed analysis was determined for the even distributed and hotspot scenarios and shown in Figure 5.7. Apparently our method does not deviate by more than 2-2.5

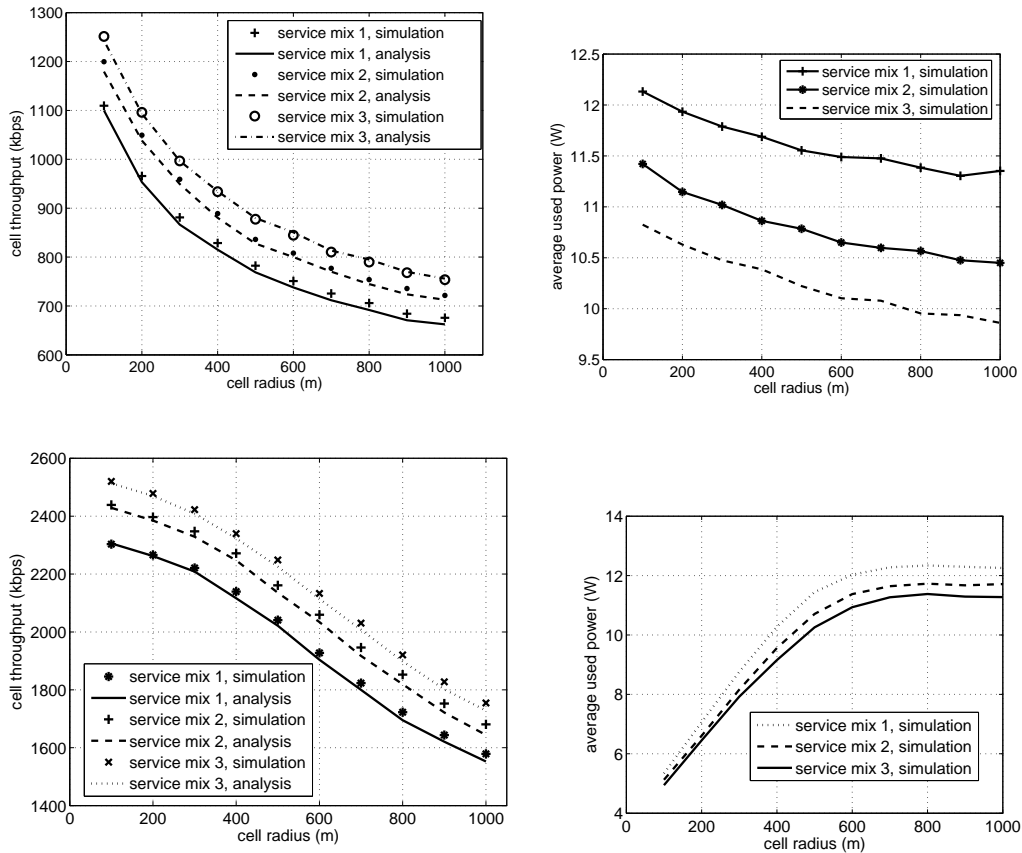


Figure 5.6: Average cell throughput (left) and used power (right), with even user distribution (top) and hotspot scenario (bottom)

percent for both scenarios. We may also conclude, that in the hotspot case there is a steady rising trend in inaccuracy, as the cell size grows, while for even distribution scenario the accuracy does not have this clear trend.

Figure 5.8 shows the same results in the concentrated scenario. As it was anticipated, the cell throughput in this case is between that of the even and hotspot scenarios. The Figure plots the average throughput of the UMTS cell as function of the cell radius in case of the three service mixes described above. In the hotspot scenario the variance of the user distribution is set to be the one-third of the cell radius, in the concentrated scenario the mean and the variance of both  $x$  and  $y$  coordinates are set to be  $\frac{2}{5}R_{\text{cell}}$  and  $\frac{R_{\text{cell}}}{5}$  respectively, where  $R_{\text{cell}}$  denotes the cell radius. Regarding the accuracy in case of concentrated scenario, it is similar to that of the hotspot scenario, with inaccuracy reaching the maximum value of 2.5%, with the increasing trend in

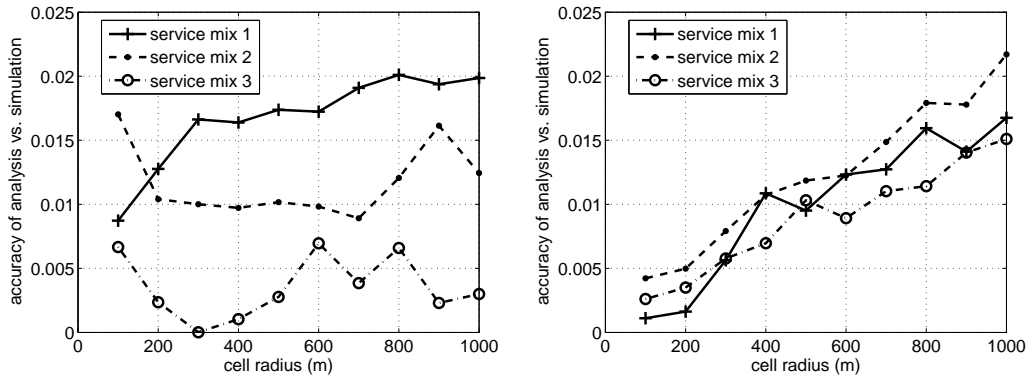


Figure 5.7: Accuracy of the numerical analysis in even distributed (left) and hotspot (right) scenario

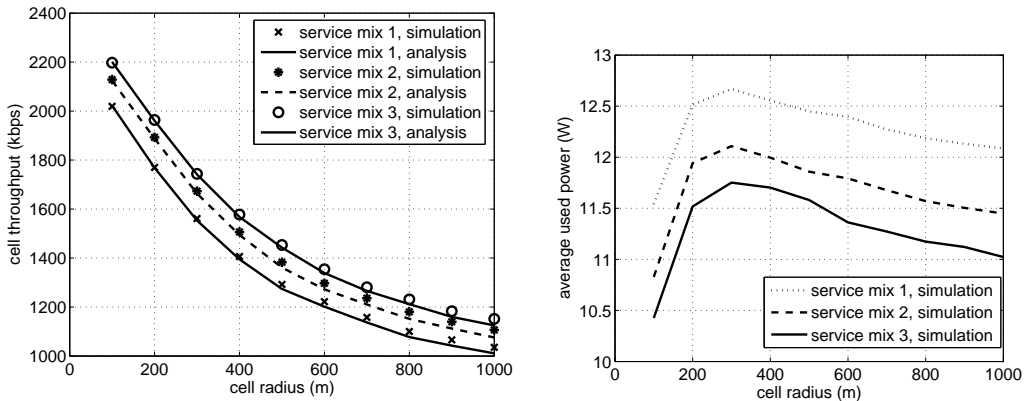


Figure 5.8: Average cell throughput (left) and used power (right), in concentrated scenario

inaccuracy as the cell size increases.

With regards to the planning of the cellular UMTS network, the following statements can be concluded. If customers are generally condensed over some distinct locations (e.g. residential blocks with rarely visited areas in between), it is worth installing base stations to the center of these locations as this result in significant increment of cell capacity. Cell size (by means of the pilot channel power) can be set to be higher: this allows the fulfillment of coverage requirements with smaller number of base stations!

With regards to the simulated values of used power, it is worth noticing, that applying the previously described simple simulation method results in different used power patterns for different user distributions. First it may seem controversial that the average used power is slightly

decreasing as cell radius increasing in even scenario. This is because of the simple simulation assumptions: generally, bigger cell means that higher power is required for a single customer, thus the remaining unused power (power of the "last" customer, with whom the total power exceeds the maximum) is higher, as a consequence the total used power is smaller. This is not the case in the hotspot scenario, as users are unlikely to be placed far from the base station.

### 5.3.2 HSDPA capacity results

This section investigates the capacity of the 3G system, if HSDPA service is deployed. Currently available 3G equipment of the market not always support the dynamic sharing of resources between conventional Release'99 traffic and HSDPA service, but requires the operator to previously configure the available resources (power, number of channelization codes) for the two types of traffic. This facilitates the capacity evaluation of HSDPA, as the effect of existing R'99 traffic can be taken into account by means of the reduced, but constant power level and spreading codes allowed for HSDPA. It is straightforward to account self interference caused by R'99 traffic with the highest power level allowed. This results in a worst case capacity estimation. Also, this method allows the evaluation of cells that do not accommodate conventional R'99 users but HSDPA traffic only. This is useful, as we anticipate the HSDPA and HSUPA services to displace conventional R'99, as all services become packet switched IP based (hence circuit switched services, along with the dedicated channel philosophy of original UMTS will slowly vanish).

Figure 5.9 shows the average cell capacity as the function of cell radius, calculated by the proposed analytical method (curves denoted by "a.") and also simulated values (dots denoted by "s."). Results were obtained for the following settings.

Five sets of radio resources (denoted by "HS set1" ... "HS set5") were defined for HSDPA, modeling situations from abundant to scarce HSDPA resources (power changing from 25 to 5 Watts along with the number of codes changing from 15 to 7). Total power of 28 Watts was considered, with constant 3 Watts pilot and signalling channel power, consequently self-interfering power was rising from 3 to 23 Watts in the five sets. Only three types of terminals were considered among the 12 possible types (type 6, 7 and 10, with assumed ratios of 0.5, 0.3 and 0.2 respectively).

It can be stated that as we anticipated it, hotspot scenario results in the highest HSDPA capacity, and even scenario provides the least HSDPA throughput. Capacity is not very sensitive

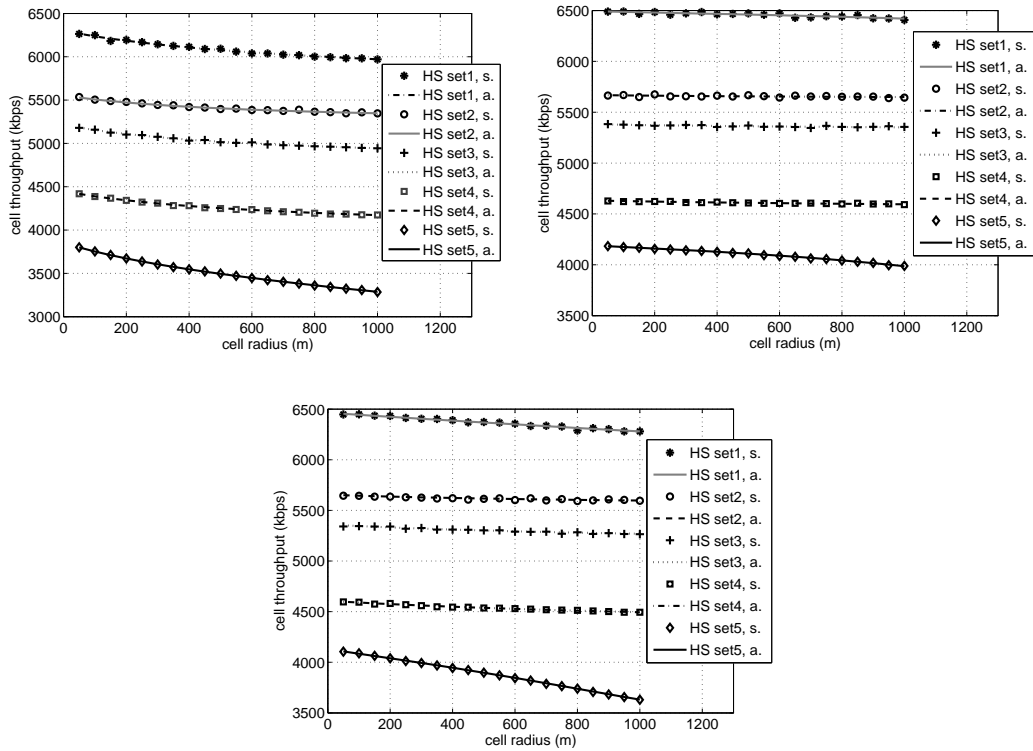


Figure 5.9: Average cell throughputs in even (left), hotspot (right) and concentrated (bottom) scenarios as function of cell size

to cell size, except in case of scarce radio resources. Average throughput does not exceed 6.5 Mbps even in case of abundant resources.

Figure 5.10 reveals average cell throughput in case of 500m cell radius, as the function of HSDPA power. Other settings are as in the previous scenario, except the number of HSDPA codes, that is also a parameter that labels the different curves. It is apparent that in case of hotspot scenario the HSDPA capacity approaches its maximum steeply as HSDPA power is increased, while in concentrated and even scenarios capacity grows less quickly. The bottom right corner of Figure 5.10 shows the average achievable throughput over the edge of the cell. This result is naturally the same for hotspot and even scenarios (as the angle of user position in polar coordinates is evenly distributed in both cases), but differs only marginally in concentrated scenario (denoted by "con" in the Figure). Although results were calculated for higher number of spreading codes, results were the same as for 8 codes. This is because the fact that over the cell edge higher rate transport formats (i.e. more spreading codes) cannot be used because of more significant intra-cell interference experienced there.

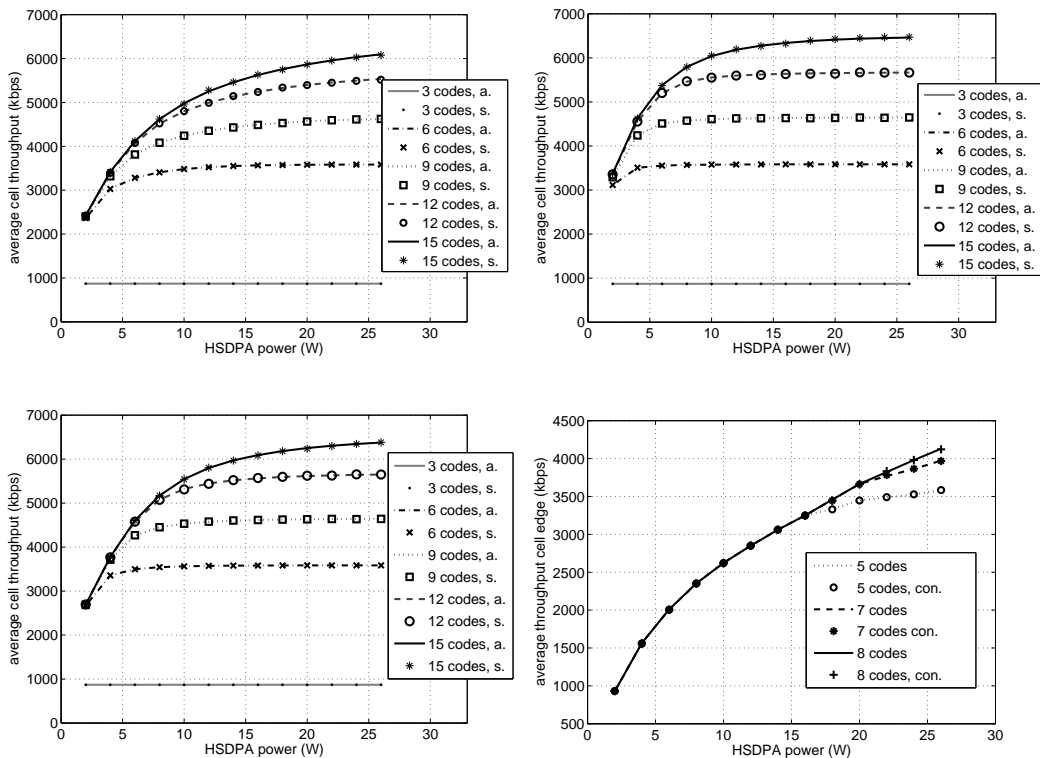


Figure 5.10: Average cell throughputs in even (top left), hotspot (top right) and concentrated (bottom left) scenarios and average throughputs over cell edge (bottom right) as function of HSDPA power

The accuracy of the analysis method is expected to be higher for these HSDPA capacity analysis cases, than for Release'99 UMTS. This is due to the fact that for Release'99 the method contain inherent approximation of user numbers of different service types by their mean values. In contrast, for the HSDPA case the analytical formulas do not contain such bias by mean approximation. The evaluation of the accuracy was conducted for all the cases plotted in Figure 5.9 and in Figure 5.10. Generally the accuracy measure does not show typical trend for cell sizes of power values and stays at very low value. The worst accuracy measure was around 0.5%, with most of the values remaining below 0.3%.

# Chapter 6

## Conclusive remarks

This dissertation is devoted to the development and presentation of analytical models and methods that allows the fast and efficient performance evaluation of cellular radio networks.

The first main group of results contain a general queueing model of mobile networks, taking into account user mobility, session duration and bursty nature of generated traffic. The queueing model is analysed using a recursive approximate solution and its accuracy is tested against simulations. The approach shown here has the advantage of enabling general distributions to model user behavior (namely phase type distributions), hence can be viewed as a generalised and conclusive version of several prior, more restrictive works. The other advantage is the capability of modelling general variable bit rate traffic sources, hence more realistic view on the network performance is achievable. The performance metrics analysed are traditionally applied for circuit switched services (namely blocking probability, radio interface utilization), however using the approach of communication sessions and bursts, connectionless, packet switched communications can also be investigated with the approach. The proposed recursive solution enables the investigation of the presented queueing model, otherwise in practical cases it would be impossible due to the resulting very large state space. This solution makes the performance analysis not only possible, but fast as well. In contrast, we get only approximate results. However, as shown by numerical results, that were obtained for the same system using the proposed algorithm and computer simulations as well, the accuracy of the approximation is reasonable. The accuracy is very good in case of moderate network load, which is the domain a cellular network should be dimensioned (namely blocking probabilities should be kept low enough). Numerical results are shown for UMTS radio interface, taking the code dimension as capacity.

Another problem investigated in the dissertation is the determination the distribution of the

residual duration of a connection that arrives to a cell after handover. This quantity is needed in order to perform correct session level analysis of a cellular system. Closed form expressions are given for general network layouts and two special, yet important network topologies are further investigated. In general case the resultant distribution is only given numerically. However, if phase type distributions are used to model customer behavior, the resultant distribution is given in analytical form and can be used directly in further analysis. Numerical results obtained via the direct method (no phase type approximation of user behavior) and results after phase type fitting to descriptive variables show very good correspondence. The resultant residual session length distributions of both methods were compared to experimental distribution obtained by computer simulation and this also fits well to the distributions obtained by calculations (not surprisingly, as here the topic is pure calculations, so matching simulation results are rather showing that the implementation of the numerical calculations and the simulator is correct).

The last topic investigated here is the problem of determining 3G radio capacity. As 3G networks are based on WCDMA radio interface, the inherent interference-limited nature of CDMA access cause that capacity, coverage and carried traffic of 3G cellular networks are strongly coupled. Moreover, multiple radio bearer types with different characteristics might be developed and used and adaptive modulation and coding is also present in HSDPA enabled networks. All these cause the radio capacity, expressed in bits per second cannot be determined easily. Therefore a calculation method is shown in this document, that defines the average capacity or throughput of the cell and shows how to calculate it, for a given average used base station output power. This analysis takes – as it is in reality – the finite transmission power as the limiting resource. Multiple bearer types are taken into account by means of the distribution of usage of different bearers, radio path loss is taken into account with appropriate propagation models, while multipath effect is considered via the use of distance dependent orthogonality factor. Numerical results show that user distribution has mayor impact on cell capacity, in a hotspot scenario, where customers tend to dwell near the base station, the carried traffic of the cell might be twice that of the capacity with evenly distributed customers. Average HSDPA capacity is also derived. The case when both Release'99 and HSDPA services are deployed on the same carrier frequency is investigated also. Here the interaction between the two is taken into account by means of the amount of used power for each services. The effect of HSDPA terminal category penetration is also taken into account, as well as the allowed code resource to HSDPA service. All numerical results regarding 3G performance are justified by results obtained from snapshot simulations.

## 6.1 Future research

One research direction within the topic of 3G analysis is straightforward: the extension of the proposed capacity model to uplink direction, including HSUPA services. The main differences and challenges of extending the downlink model are because of the fact that intra-cell interference powers are attenuated individually for each user (whereas in downlink, the attenuation of interferers' signals and useful power was the same) and inter-cell interference is the sum of powers of randomly placed users (while in downlink, outer interference could be well modelled by a given power level from a fixed neighboring Node B location). Moreover, the level of outer terminal powers is affected by the power level of terminals in the examined cell. Intuitively it seems that the method of writing the expression of average terminal power and define the ratio of different radio bearers may work. The problem of interacting powers in neighboring cells may be handled by either solving the system of average power equations for the whole network under examination, or to use an iterative method. The latter would start with solving the power equations for a given cell, supposing no outer interference, than solving the equations for a neighboring cell, with the interfering powers calculated in the previous step for the previous cell, and continuing iteratively with substituting updated power values until convergence. To model HSUPA service, the first approach could be – as HSUPA uses power controlled dedicated channels as Release'99 UMTS – to simply consider HSUPA as specific radio bearer types. However, HSUPA scheduling enables not only the change of transmit power, but the adaptation of transport format to changing radio environment. Hence, if the level of interference does not allow a HSUPA connection with a given bitrate, it may switch to another, lower bitrate transport format.

Another straightforward step toward more detailed evaluation of joint Release'99-HSDPA performance is the computation of not only the average, but the distribution of used powers and customer numbers. One idea to calculate the exact distribution of the power allocated to a user, from (5.9). However, by examining the expression we may conclude that it is computationally unattractive. First, the total used power is also present in the equation (that is dependent on the distribution we want to determine). On the other hand, the other to own path loss ratios are summed in the equation, so the convolution of multiple distributions are required. These can be overcome if we suppose that HSDPA traffic is always present, then the random term  $P_{\text{inst}}^0$  should be replaced with the constant base station power  $P_0^0$ . On the other hand, the other to own cell path loss ratio could be approximated by some distribution. In the literature sometimes it is approximated by lognormal distribution. In this case, the distribution of the sum of these factors

will be again well approximated by lognormal. The distribution of the orthogonality factor can be easily determined by variable transformation. This yields that the random terms in the equation will have closed form distributions, their sum has to be calculated using numerical convolution. The other option to use PH fitting to the distribution of the sum other to own cell path loss ratios and the orthogonality factor, in this case the user power distribution will be in hand as having also a PH distribution. after having the distribution of transmitted power to a type  $k$  user, this will allow the following analysis: determination of the sum Release'99 power distribution under any traffic mix and thus the determination of HSDPA power distribution. Having this will allow the determination of the distribution of HSDPA signal to interference ratio from (5.22). As the denominator is analysed in the previous step, the last (but not trivial) task is to compute the distribution of the fraction of the two resultant random quantities. From the random SIR the mapped random CQI and achievable bitrate follows. This computation will give much deeper insight into the performance of 3G cellular systems, moreover, it allows the development of an elaborate queueing model. The skeleton of this queueing modelling could be: the flow of sessions using any Release'99 radio bearer is supposed to be Poissonian and the used power distribution is given previously. Then the occupied total power can be analysed by for example the "stochastic knapsack with continuous sizes" model [102] to determine the cdf of total occupied power (and session blocking performance of Release'99 connections). If this continuous handling proves to be computationally infeasible, than the discretization of the power levels and assigning discrete probabilities to these result in well known loss queueing model. The HSDPA performance is the analysed as following: supposing that there is a scheduler that provides fair share of the radio capacity (in terms of bitrate achieved) in all possible states of residual radio resource (power and codes) the HSDPA is analysed as a processor sharing system, with the average HSDPA throughput determined for each state. The overall performance is then calculated by summing up all the results weighted by the state probabilities.

It is inevitable to expand the capacity analysis shown in this dissertation to the recently standardized radio interface of 3GPP LTE (Long Term Evolution, this system is also often referred as Enhanced UTRA). The method shown here is applicable more or less directly, if there are results on the usable transport formats (hence transmission rates) as function of the signal to interference ratio. Somehow this analysis will be simpler than that of 3G, because of the fact that users are separated in time and frequency domain, self-interference will not occur in perfect LTE system. The detailed approach would require the investigation of network topologies with different frequency allocations (different bandwidths might be used in different cells, these can

be overlapping as well). In LTE it is an essential requirement from the industry, that frequency reuse of 1 should be able to be used (same bands in neighboring cells). This can be achieved by means of intelligent scheduling, that avoids the allocation of the same physical resource (carrier frequency and timeslot) for users dwell near cell borders, as the required high power would cause high interference. This mechanism – without the detailed knowledge of such operation – should also be taken into account when investigating neighboring cell interference issues in LTE.

# Bibliography

- [1] 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*, 35(3):77–92, August 1986.
- [2] M. Ajmone Marsan, S. Marano, C. Mastroianni, and M. Meo. Performance analysis of cellular mobile communication networks supporting multimedia services. *Mobile Networks and Applications*, 5(4):167–177, 2000.
- [3] Moshe Sidi and David Starobinski. New call blocking versus handoff blocking in cellular networks. *Wireless Networks*, (3):15–27, 1997.
- [4] Yi-Bing Lin, Li-Fang Chang, and A. Noerpel. Modeling hierarchical microcell/macrocell PCS architecture. In *IEEE International Conference on Communications, ICC 1995, Seattle*, volume 1, pages 405–409, 1995.
- [5] A. L. Wijesinha, S. P. Kumar, and D. P. Sidhu. Handover and new call blocking performance with dynamic single-channel assignment in linear cellular arrays. *Wireless Networks*, 6(2):121–130, 2000.
- [6] Mahmoud Naghshineh and Anthony S. Acampora. QoS provisioning in micro-cellular networks supporting multiple classes of traffic. *Wireless Networks*, 2(3):195–203, 1996.
- [7] Taekyoung Kwon, Yanghee Choi, Chatschik Bisdikian, and Mahmoud Naghshineh. QoS Provisioning in Wireless/Mobile Multimedia Networks Using an Adaptive Framework. *Wireless Networks*, (9):51–59.
- [8] Y-B. Lin and I. Chlamtac. A model with generalized holding and cell residence times for evaluating handoff rates and channel occupancy times in PCS networks. *International Journal of Wireless Information Networks*, 4(3):163–171, July 1997.

- [9] Wei Feng and Masashi Kowada. Performance analysis of wireless mobile networks with queueing priority and guard channels. *International Transactions in Operational Research*, 15(4):481–508, July 2008.
- [10] Li Wei, Chen Hang, and D.P. Agrawal. Performance analysis of handoff schemes with preemptive and nonpreemptive channel borrowing in integrated wireless cellular networks. *Wireless Communications, IEEE Transactions on*, 5(3):1222–1233, May 2005.
- [11] Yan Zhang, Yang Xiao, and Hsiao Chen. Queueing analysis for OFDM subcarrier allocation in broadband wireless multiservice networks. *Wireless Communications, IEEE Transactions on*, 7(10):3951–3961, October 2008.
- [12] Weiwei Wu and T. Sakurai. Capacity of reuse partitioning schemes for OFDMA wireless data networks. In *Proc. of Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on*, pages 2240–2244, September 2009.
- [13] Performance analysis of cellular mobile telephone networks with hybrid channel assignment scheme. In Wuyi Yue and Yutaka Matsumoto, editors, *Performance Analysis of Multichannel and Multi-Traffic on Wireless Communication Networks*, pages 245–263. Springer-Verlag Berlin, 2010.
- [14] Lan Wang, Geyong Min, Demetres Kouvatsos, and Xiangxiang Zuo. Modelling and analysis of a dynamic guard channel handover scheme with heterogeneous call arrival processes. In Demetres Kouvatsos, editor, *Network Performance Engineering*, volume 5233 of *Lecture Notes in Computer Science*, pages 665–681. Springer Berlin / Heidelberg, 2011.
- [15] Anum L. Enlil Corral-Ruiz, A. Cruz-Perez, and Genaro Hernandez-Valdez. Coxian Distribution Modeling for the Generalized and Unified Teletraffic Analysis of Mobile Cellular Networks. In *Proc. of 7th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2010)*, September 2010.
- [16] H. Zeng and I. Chlamtac. Adaptive guard channel allocation and blocking probability estimation in PCS networks. *Computer Networks*, 43(2):163–176, 2003.
- [17] K. S. Gilhousen, I. M. Jakobs, 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*, 40(2):303–397, May 1991.

- [18] A. M. Viterbi and A. J. Viterbi. Erlang capacity of a power controlled CDMA system. *IEEE Journal on Selected Areas in Communications*, 11(6), August 1993.
- [19] K. Sipila, K. C. Honkasalo, J. Laiho-Steffens, and A. Wacker. Estimation of capacity and required transmission power of WCDMA downlink based on a downlink pole equation. In *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*, volume 2, pages 1002–1005, 2000.
- [20] Kimmo Hiltunen and Riccardo De Bernardi. WCDMA downlink capacity estimation. In *Proc. of IEEE Vehicular Technology Conference, VTC 2000 Spring*, volume 2, pages 992–996, May 2000.
- [21] W. Choi and J. Y. Kim. Forward-link capacity of a DS/CDMA system with mixed multi-rate sources. *IEEE Transactions on Vehicular Technology*, 50(3):737–749, May 2001.
- [22] A. J. Viterbi. *CDMA: Principles of Spread Spectrum Communication*. Addison-Wesley Wireless Communications Series, 1995.
- [23] Qinqing Zhang. UMTS air interface voice/data capacity-part 2: forward link analysis. In *Proc. of IEEE Vehicular Technology Conference, VTC 2001 Spring*, volume 4, pages 2730–2734, May 2001.
- [24] Andreas Mader and Dirk Staehle. An analytic model for deriving the Node-B transmit power in heterogeneous UMTS networks. In *Proc. of IEEE Vehicular Technology Conference, VTC 2004 Spring*, May 2004.
- [25] Andreas Mader and Dirk Staehle. Analytic modelling of the WCDMA downlink capacity in multi-service environments. Technical report, University of Wurzburg, 2004.
- [26] Chie Dou. The Maximum Available Radio Resource of a WCDMA Downlink. *IEICE Transactions on Communications*, E88-B(11):4309–4316, November 2005.
- [27] R. Litjens L. van den Berg and J. Laverman. HSDPA flow level performance: the impact of key system and traffic aspects. In *Proceedings of Seventh IEEE/ACM MSWiM*, October 2004.
- [28] M. Wrulich, W. Weiler, and M. Rupp. HSDPA performance in a mixed traffic network. In *Proc. of IEEE Vehicular Technology Conference, VTC 2008 Spring*, pages 2056–5060, May 2008.

- [29] M. Assaad and D. Zeghlache. On the Capacity of HSDPA. In *Proc. of IEEE Global Telecommunications Conference, GLOBECOM'03*, volume 1, pages 60–64, 2003.
- [30] K. I. Pedersen, T. F. Lootsma, M. Stottrup, F. Frederiksen, T. E. Kolding, and P. E. Mogenssen. Network Performance of Mixed Traffic on High Speed Downlink Packet Access and Dedicated Channels in WCDMA. In *Proc. of IEEE Vehicular Technology Conference, VTC 2004 Fall*, pages 4496–4500.
- [31] E. Altman, T. Chahed, and S.E. Elayoubi. Joint uplink and downlink capacity considerations in admission control in multiservice CDMA/HSDPA systems. In *Proc. of ACM 2nd international conference on Performance evaluation methodologies and tools*, 2007.
- [32] Qualcomm Engineering Services Group. Air Interface Cell Capacity of WCDMA Systems. Technical report, QUALCOMM Incorporated, 2007.
- [33] Christophe Chevallier et. all. *WCDMA (UMTS) deployment handbook. Planning and Optimization Aspects*. John Wiley & Sons inc., 2006.
- [34] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, june 1995.
- [35] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. In *ACM SIGCOMM Proceedings of the conference on Communications architectures, protocols and applications*, pages 257 – 268, 1994.
- [36] E. Chlebus and W. Ludwin. Is handoff traffic really Poissonian. In *Proc. of IEEE ICUPC'95*, pages 348–353, 1995.
- [37] M. Rajaratnam and F. Takawira. Handoff traffic modelling in cellular networks. In *Proc. of IEEE GLOBECOM'97, Global Telecommunications Conference*, pages 131–137, 1997.
- [38] P. V. Orlik and S. S. Rappaport. On the handoff arrival process in cellular communications. *Wireless Networks*, 7(2):147–157, March/April 2001.
- [39] M. M. Zonoozi, P. Dassanayake, and M. Faulkner. Mobility modelling and channel holding time distribution in cellular mobile communication systems. In *Proc. of IEEE GLOBECOM'95, Global Telecommunications Conference*, volume 1, pages 12 –16, 1995.

- [40] M. Zonoozi and P. Dassanayake. User mobility modeling and characterization of mobility patterns. *IEEE Journal on Selected Areas in Communications*, 15(7):1239–1252, September 1997.
- [41] L. Kleinrock. *Queueing systems. Volume I: Theory*. John Wiley & Sons inc., 1975.
- [42] M. Neuts. *Probability distributions of Phase Type*, volume Liber Amicorum Prof. Emeritus H. Florin, pages 173–206. University of Leuven, 1975.
- [43] S. Asmussen. Phase-type distributions and related point processes: fitting and recent advances. In *Proceedings of the First International Conference on Matrix Analytic Methods in Stochastic Models*, volume 1, pages 137–149, 1997.
- [44] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distribution via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.
- [45] A. Riska, V. Diev, and E. Smirni. Efficient fitting of long-tailed data sets into phase-type distributions. *Performance Evaluation*, 55:147–164, 2004.
- [46] A. Riska, V. Diev, and E. Smirni. Efficient fitting of long-tailed data sets into hyperexponential distributions. In *Proceedings of the IEEE Internet Performance Symposium*, 2002.
- [47] A. Horváth and M. Telek. Approximating heavy tailed behaviour with phase type distributions. In *Proceedings of the 3rd International Conference on Matrix-Analytic Methods in Stochastic Models*, pages 191–214, 2000.
- [48] A. Bobbio and M. Telek. A benchmark for PH estimation algorithms: results for Acyclic-PH. *Stochastic models*, 10:661–677, 1994.
- [49] Axel Thummler, Peter Buchholz, and Miklos Telek. A Novel Approach for Phase-Type Fitting with the EM Algorithm. *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, 3(3):245–258, Suly-September 2006.
- [50] S. S. Rappaport. Blocking, hand-off and traffic performance for cellular communication systems with mixed platforms. *IEE Proceedings of Communications, Speech and Vision*, 140(5):389–401, October 1993.

- [51] P. V. Orlik and S. S. Rappaport. Traffic performance and mobility modeling of cellular communications with mixed platforms and highly variable mobilities. *Proceedings of the IEEE*, 86(7):1464–1479, July 1998.
- [52] P. V. Orlik and S. S. Rappaport. A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions. *IEEE Journal on Selected Areas in Communications*, 16(5):788–803, June 1998.
- [53] Y. Fang and I. Chlamtac. Teletraffic analysis and mobility modeling of PCS networks. *IEEE Transactions on Communications*, 47(7):1062–1072, July 1999.
- [54] Y. Fang. Hyper-Erlang distributions and traffic modeling in wireless and mobile networks. In *Proceedings of the IEEE WCNC'99, Wireless Communications and Networking Conference*, volume 1, pages 398–402, 1999.
- [55] Y. Fang. Hyper-erlang distribution model and its application in wireless mobile networks. *Wireless Networks*, 7(3):211–219, May 2001.
- [56] K. Yeo and C-H. Jun. Modeling and analysis of hierarchical cellular networks with general distributions of call and cell residence times. *IEEE Transactions on Vehicular Technology*, 51(6):1361–1374, November 2002.
- [57] Y-B. Lin and I. Chlamtac. Effects of erlang call holding times on PCS call completion. In *Proc. of IEEE INFOCOM'97, Sixteenth Annual Joint Conference of the IEEE Computer and Communication Societies*, 1997.
- [58] A. Jayasuriya, D. Green, and J. Asenstorfer. Modelling service time distribution in cellular networks using phase type service distributions. In *Proceedings of IEEE ICC'01 International Conference on Communications*, pages 440–444, 2001.
- [59] A. S. Alfa and W. Li. A homogeneous PCS network with Markov call arrival process and phase type cell residence time. *Wireless Networks*, 8(6):597–605, November 2002.
- [60] V. Pla and V. Casares-Giner. Analytical-numerical study of the handoff area sojourn time. 2002.

- [61] H. Zeng and I. Chlamtac. Handoff traffic distribution in cellular networks. In *Proceedings of IEEE WCNC'99 Wireless Communications and Networking Conference*, volume 1, pages 413–417, 1999.
- [62] H. Zeng, Y. Fang, and I. Chlamtac. Call blocking performance study for PCS networks under more realistic mobility assumptions. *Telecommunication Systems*, 19(2):125–146, 2002.
- [63] C. Jedrzycki and V. C. M. Leung. Probability distribution of channel holding time in cellular telephony systems. In *Proceedings of IEEE VTC'96 Vehicular Technology Conference*, pages 247–251, 1996.
- [64] S. Thajchayapong and J. M. Peha. Mobility patterns in microcellular wireless networks. In *Proceedings of IEEE WCNC'03 Wireless Communications and Networking Conference*, 2003.
- [65] F. Chang and W. Feng. Modeling player session times of on-line games. In *Proceedings of the 2nd Workshop on Network and System Support for Games*, pages 23–26, 2003.
- [66] T. Henderson and S. Bhatti. Modelling user behavior in networked games. In *Proceedings of ACM Multimedia*, pages 212–220, 2001.
- [67] J. Jordán and F. Barceló. Statistical modelling of transmission holding time in PAMR systems. In *Proceedings of the IEEE Globecom'97 Global Telecommunications Conference*, volume 1, pages 121–125, 1997.
- [68] J. Jordán, F. Barceló, and J. Paradells. Voice holding time distribution in trunked PAMR systems. In *Proceedings of the IEEE VTC'97 Vehicular Technology Conference*, volume 1, pages 436–440, 1997.
- [69] J. Jordán and F. Barceló. Channel holding time distribution in cellular telephony. In *Proceedings of Wireless'97 9th International Conference on Wireless Communications*, pages 125–134, 1997.
- [70] J. Jordán and F. Barceló. Channel holding time distribution in cellular telephony. *IEE Electronics Letters*, 34(2):146–147, 1998.

- [71] J. Jordán and F. Barceló. Channel holding time distribution in public telephony systems (PAMR and PCS). *IEEE Transactions on Vehicular Technology*, 49(5):1615–1625, September 2000.
- [72] Ioannis Z. Koukoutsidis, Petros I. Papaioannou, and Michael E. Theologou. Effect of cell residence time variance on the performance of an advanced paging algorithm. *CoRR*, abs/0904.0771, 2009.
- [73] R. J. Gibbens and P. J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems: Theory and Applications*, 9(1-2):17–28, October 1991.
- [74] F. P. Kelly. Effective bandwidths at multi-type queues. *Queueing Systems: Theory and Applications*, 9(1-2):5–16, October 1991.
- [75] George Kesidis, Jean Walrand, and Cheng-Shang Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE-ACM Transactions on Networking*, 1(4):424–428, 1993.
- [76] A. I. Elwalid and D. Mitra. Effective bandwidth of general markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking*, 1(3):329–343, 1993.
- [77] F. Adachi, M. Sawahashi, and K. Okawa. Tree-structured generation of orthogonal spreading codes with different length for forward link of DS-CDMA mobile radio. *Electronics Letters*, 33(1):27–28, January 1997.
- [78] A.N. Rouskas and D.N. Skoutas. OVFS codes assignment and reassignment at the forward link of W-CDMA 3G systems. In *Proceedings of the IEEE PIMRC'02 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, volume 5, pages 2404–2408, 2002.
- [79] H. Cam. Nonblocking OVFS codes and enhancing network capacity for 3G wireless and beyond systems. *Computer Communications, spec. issue on 3G Wireless and Beyond for Computer Communications*, 26(17):1907–1917, November 2003.
- [80] S. Choi and Kang G. Shin. An uplink CDMA system architecture with diverse QoS guarantees for heterogeneous traffic. *IEEE/ACM Transactions on Networking*, 7(5):616–628, October 1999.

- [81] S. J. Lee, T. S. Kim, and D. K. Sung. Bit-error probabilities of multicode direct-sequence spread-spectrum multiple-access systems. *IEEE Transactions on Communications*, 49(1):31–34, January 2001.
- [82] D. I. Kim and V. K. Bhargava. Performance of multidimensional multicode DS-CDMA using code diversity and error detection. *IEEE Transactions on Communications*, 49(5):875–887, May 2001.
- [83] P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou. Models for packet switching of variable-bit-rate video sources. *IEEE Journal on Selected Areas in Communications*, 7(5):865–869, June 1989.
- [84] D. Gan and S. McKenzie. Source modelling for B-ISDN networks with ATM switching. In *Proc of the IEE colloquium on Multimedia Communications Systems*, March 1994.
- [85] J. F. Frigon, H. C. B. Chan, and V. C. M. Leung. A variable bit rate resource allocation algorithm for wireless ATM. In *Proc. of IEEE GLOBECOM'99, Global Telecommunications Conference*, volume 5, pages 2673–2677, 1999.
- [86] A. La Corte, A. Lombardo, and G. Schembra. Modeling superposition of ON-OFF correlated traffic sources in multimedia applications. In *Proceedings of IEEE INFOCOM'95, Fourteenth Annual Joint Conference of the IEEE Computer and Communication Societies*, volume 3, pages 993–1000, 1995.
- [87] A. La Corte, A. Lombardo, and G. Schembra. Analysis of packet loss in a continuous-time finite-buffer queue with multimedia traffic streams. *International Journal of Communication Systems*, 10(2):329–343, 1997.
- [88] S. Galmes and R. Puigjaner. On the capabilities of on-off models to capture arbitrary ATM sources. In *Proceedings of the IEEE INFOCOM'98 Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 3, pages 1333–1340, 1998.
- [89] N. K. Shankaranarayanan, Z. Jiang, and P. Mishra. User-perceived performance of web-browsing and interactive data applications in TDMA packet wireless networks. In *Proceedings of IEEE MMT'2000 Multiaccess, Mobility and Teletraffic for Wireless Communications*, 2000.

- [90] Z. Jiang, L. Chang, and N. K. Shankaranarayanan. Providing multiple service classes for bursty data traffic in cellular networks. In *Proceedings of IEEE INFOCOM'2000, Nineteenth Annual Joint Conference of the IEEE Computer and Communication Societies*, 2000.
- [91] M. Molina, P. Castelli, and G. Foddis. Web traffic modeling exploiting TCP connections' temporal clustering through HTML REDUCE. *IEEE Network Magazine*, 14(3):46–55, May/June 2000.
- [92] H.-K. Choi and J. O. Limb. A behavioral model of web traffic. In *Proceeding of ICNP'99 Seventh Annual International Conference on Network Protocols*, pages 327–334, 1999.
- [93] Shriram Sarvotham, Rudolf Riedi, and Richard Baraniuk. Network and user driven alpha-beta on-off source model for network traffic. *Computer Networks*, 48:335–350, 2005.
- [94] Xiangjun He and Tigang Jiang. P-persistent CSMA Protocol Simulation Based on ON/OFF Source Model in Cognitive Radio Network. In *Proc. of Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, pages 1–3, September 2010.
- [95] N. K. Shankaranarayanan, A. Rastogi, and Z. Jiang. Performance of a wireless data network with mixed interactive user workloads. In *Proceedings of IEEE ICC'02 International Conference on Communications*, 2002.
- [96] P. Barford and M. Crovella. Generating representative web workloads for network and server performance evaluation. In *Proceedings of ACM Sigmetrics'98 International Conference on Measurement and Modeling of Computer Systems*, pages 151–160, 1998.
- [97] D. Staehle, K. Leibnitz, and P. Tran-Gia. Source traffic modeling of wireless applications. Technical report, University of Wurzburg, 2000.
- [98] R. W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30:223–231, 1982.
- [99] F. Baskett, K. M. Chandy, R. R. Muntz, and F. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22:248–260, 1975.
- [100] X. Chao, M. Miyazawa, R.F. Serfozo, and H. Takada. Markov network processes with product form stationary distributions. *Queueing Systems*, 28(4).

- [101] Erol Gelenbe and Guy Pujolle. *Introduction to Queueing Networks*. John Wiley & Sons inc., 1998.
- [102] Keith W. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag London Limited, 1995.
- [103] J. Kaufman. Blocking in a shared resource environment. *IEEE Transactions on Communications*, Com-29(10):1474–1481, October 1981.
- [104] J. W. Roberts. *A service system with heterogeneous user requirements*, volume Performance of Data Communication Systems and their Applications, pages 423–431. North-Holland-Elsevier Science Publishers, 1981.
- [105] Y. C. Tseng, C. M. Chao, and S. L. Wu. Code placement and replacement strategies for wideband CDMA OVSF code tree management. In *Proceedings of IEEE Globecom 2001*, volume 1, pages 562–566, 2001.
- [106] Y. C. Tseng and C. M. Chao. Code placement and replacement strategies for wideband CDMA OVSF code tree management. *IEEE Transactions on Mobile Computing*, 1(4):293–302, October-December 2002.
- [107] R. G. Cheng and P. Lin. OVSF code channel assignment for IMT-2000. In *Proceedings of IEEE Vehicular Technology Conference, VTC 2000, Tokyo*, volume 3, pages 2188–2192, 2000.
- [108] A. Aggarwal and D. Gupta. OVSF code allocation in CDMA based wireless ad hoc networks. Technical report, Dept. of Computer Sc. and Eng., Indian Institute of Technology, Kanpur, April 2003.
- [109] Anum L. Enlil Corral-Ruiz, Andres Rico-Paez, Felipe A. Cruz-Perez, and Genaro Hernandez-Valdez. On the functional relationship between channel holding time and cell dwell time in mobile cellular networks. In *Proc. of IEEE GLOBECOM 2010*, pages 1–6. IEEE, 2010.
- [110] N. B. Mehta, L. J. Greenstein, T. M. Willis, and Z. Kostic. Analysis and results for the orthogonality factor in WCDMA downlinks. *IEEE Transactions on Wireless Communications*, 2(6), November 2003.

- [111] K. I. Pedersen and P. E. Mogensen. The downlink orthogonality factors influence on WCDMA system performance. In *Proceedings of the 56th IEEE Vehicular Technology Conference, VTC 2002-Fall*, volume 4, pages 2061–2065, 2002.
- [112] N. B. Mehta, A. F. Molisch, and L. J. Greenstein. Orthogonality factor in WCDMA downlinks in urban macrocellular environments. In *Proceedings of the IEEE Global Telecommunications Conference, GLOBECOM'05*, volume 6, 2005.
- [113] N. B. Mehta, A. F. Molisch, and L. J. Greenstein. Macrocell-Wide behavior of the Orthogonality Factor in WCDMA Downlinks. *IEEE Transactions on Wireless Communications*, 5(12), December 2006.
- [114] Pablo Jose Ameigeiras Gutierrez. *Packet Scheduling and Quality of Service in HSDPA*. PhD thesis, Aalborg University, Department of Communication Theory, October 2003.
- [115] F. Brouwer, I. de Bruin, J.C. Silva, N. Souto, F. Cercas, and A. Correia. Usage of link-level performance indicators for HSDPA network-level simulations in E-UMTS. *Proceedings of IEEE ISSSTA'04, Sydney*, 2004.
- [116] Motorola and Ericsson. Revised CQI proposal. 3GPP TSG-RAN-WG1 HSDPA, R1-02-0675, April 9-12, 2002.

# Appendix A

## Scientific publications of the author

### Journal papers

- [J1] Fazekas P., Imre S., Telek M., "Performance evaluation of multimedia services in cellular networks", *Simulation-Transactions of the Society for Modeling and Simulation International*, Vol 78., Issue 4., pp 268-277, April 2002;
- [J2] P. Fazekas, S. Imre and M. Telek, "Modeling and Analysis of Broadband Cellular Networks with Multimedia Connections", *Telecommunication Systems*, Vol 19., Issues 3-4, pp. 263-288, March-April 2002;
- [J3] Fazekas Péter, "Mobicom '97 konferencia - az ATM és IP jegyében", *Magyar Távközlés*, VIII. évfolyam, 10. szám, 48-51. old., október, 1997;
- [J4] P. Fazekas and S. Imre, "Traffic Analysis of Multimedia Services in Broadband Cellular Networks", *Lecture Notes in Computer Science 2093*, pp. 296-306, 2001, (Originally: *Proc. of Networking-ICN 2000*, Vol 1, Colmar, France, July 9-13, 2001)
- [J5] Fazekas P., "Többvívős kódosztásos rendszerek", *Magyar Távközlés*, IX. évfolyam, 10. szám, 21-25. old., október, 1998;
- [J6] Fazekas Péter, Imre Sándor, Pap László, Schulcz Róbert, Szabó Sándor, "3. generációs mobil rendszerek hívásengedélyezési módszerei", *Magyar Távközlés*, X. évfolyam, 1. szám, 18-21. old., január, 1999;
- [J7] Fazekas Péter, Imre Sándor, Jeney Gábor, Pap László, Schulcz Róbert, Szabó Sándor, "Nagysebességű vezeték nélküli hálózatok - a közeljövő technológiái." *HÍRADÁSTECHNIKA LXIV: pp. 34-42. (2009)*

## Book chapters

[B1] Péter Fazekas, "UMTS: 3rd Generation Cellular Mobile Radio System." *In: Katalin Tarnay, Gusztáv Adamis, Tibor Dulai (editors) Advanced Communication Protocol Technologies: Solutions, Methods, and Applications.* Hershey ; New York: IGI Global, Information Science Reference, 2011. pp. 134-154. (ISBN: ISBN 978-1-60960-732-6)

## International conference papers

[C1] P. Fazekas, L. Pap, S. Imre, "Joint Detection in Spread Spectrum Mobile Communication Systems", *Proc. of Poster '98 International Workshop on Scientific Electrical Engineering*, Prague, Czech Republic, May 27-29

[C2] P. Fazekas, S. Imre, "Modeling a Virtual Connection Tree based mobile ATM network", *Proc. of the IEEE VTC'99 Fall*, pp. 539-544, Amsterdam, Netherlands, Sep 19-22, 1999

[C3] P. Fazekas, "Support of Handover in Mobile ATM Networks", *Proc. of EUNICE'99 Fifth Open European Summer School*, pp. 107-112, Barcelona, Spain, Sep 1-3, 1999;

[C4] P. Fazekas, S. Imre, M. Telek, "Analysis of broadband cellular networks with variable bit-rate connections", *Proc. of EUNICE 2000 Sixth Open European Summer School*, pp. 211-217, Enschede, The Netherlands, Sep. 13-15, 2000

[C5] P. Fazekas, S. Imre: "Performance Evaluation of Cellular Networks with Multiple Multimedia Service Classes", *Proc of 6th International Conference on Telecommunications, CON-TEL2001*, pp. 159-167, Zagreb, Croatia, June 13-15, 2001

[C6] P. Fazekas, L. Pap, "On the residual session length of multimedia connections in broadband cellular networks", *Proc. of IST Mobile & Wireless Telecommunications Summit*, pp. 217-221, Thessaloniki, Greece, June 17-19, 2002

[C7] P. Fazekas, "Calculating the Session Length Distribution of Handover Customers in Multimedia Cellular Networks", *Proc. of IST 2005 International Symposium on Telecommunications*, vol. 2., pp. 767-772, Shiraz, Iran, September 10-12, 2005

[C8] Lucio Ferreira, Jordi Perez-Romero, Velio Tralli, Peter Fazekas, Miquel Oliver, Stefan Lindskog, Ramón Agustí, "QoS Provision in Wireless Networks: Mobility, Security, and Radio Resource Management: An Overview.", Presented by Peter Fazekas, at *IEEE International Conference on Communications (ICC 2006)*, Istanbul, Turkey, June 11-15, 2006 (2006)

[C9] Péter Fazekas, "Analysis of packet level QoS in wireless data networks." *3rd Symposium*

on *Wireless Communication Systems, ISWCS 2006*, Valencia, Spain, 5-8 September, 2006

[C10] Zoltán Faigl, Péter Fazekas, Stefan Lindskog, Anna Brunstrom, "Performance Analysis of IPsec in Mobile IPv6 Scenarios." *In: Proceedings of the 16th IST Mobile and Wireless Communications Summit 2007*. Budapest, Magyarország, 2007.07.01-2007.07.05. pp. 1-5. Paper 4299278. (ISBN: 963-8111-66-6)

[C11] Zoltán Faigl, Péter Fazekas, Stefan Lindskog, Anna Brunstrom, "Analytical Analysis of the Performance Overheads of IPsec in MIPv6 Scenarios." *In: Frigyes István, János Bitó, Péter Bakki (editors.) Advances in Mobile and Wireless Communications: Views of the 16th IST Mobile and Wireless Communication Summit*. Berlin ; Heidelberg: Springer-Verlag, 2008. pp. 365-385. (Lecture Notes in Electrical Engineering; 16.) (ISBN: 978-3-540-79040-2)

[C12] Fazekas Péter, "Achievable Cell throughput in 3G Systems." *In: Proc. of Eurosis ISC 2010 Conference*. Budapest, Magyarország, 2010.06.05-2010.06.07. pp. 203-208.

[C13] Fazekas Péter, Jakab Tivadar, Sipos Attila, "Egyszerű modellszámítások országos kiterjedésű hálózatra FLEXPLANET alapokon." *In: HTE Infokom 2010: Intelligens infrastruktúrák és alkalmazások*. Siófok, Magyarország, 2010.10.27-2010.10.29. Budapest: pp. 1-6.

[C14] G Auer, I Gódor, L Hévízi, M Imran, J Malmudin, P Fazekas, Gergely Biczók, H Holtkamp, D Zeller, O Blume, R Tafazolli, "Enablers for Energy Efficient Wireless Networks.", *In: Green Wireless Communications and Networks Workshop - GreeNet, Proc. IEEE VTC Fall 2010*. Ottawa, Canada, 2010.09.06-2010.09.08. (IEEE) IEEE Press, pp. 1-5. Paper 1.

[C15] G Auer, I Godor, L Hevizi, M A Imran, J Malmudin, P Fazekas, G Biczok, D Zeller, O Blume, R Tafazolli, "The EARTH Project: Towards Energy Efficient Wireless Networks." *In: Paul Cunningham, Miriam Cunningham (editors) Future Network & Mobile Summit: Conference Proceedings*. Firenze, Italy, 2010.06.16-2010.06.18. Firenze: pp. 1-5. Paper 226. (ISBN: 978-1-905824-16-8)

[C16] Törös István, Fazekas Péter, "An Algorithm for Automatic Base Station Placement in Cellular Network Deployment." *In: EUNICE 2010, LNCS 6164 Networked Services and Applications - Engineering, Control and Management*. Trondheim, Norway, 2010.06.28-2010.06.30. pp. 21-30. Paper 3.

[C17] Törös István, Fazekas Péter, "Automatic Base Station Deployment Algorithm in Next Generation Cellular Networks." *In: AccessNets 2010: 5th International ICST Conference on Access Networks*. Budapest, Hungary, 2010.11.03-2010.11.05. Budapest: pp. 1-14. Paper 2. (ISBN: 978-963-9995-09-3)

[C18] Albert Mráz, Péter Fazekas, "Effect of Imperfect Channel Estimation on LTE MU-MIMO

Performance." In: *Paul Cunningham, Miriam Cunningham (editor) Future Network and Mobile Summit 2011 Conference Proceedings*. Warsaw, Poland, 2011.06.15-2011.06.17. pp. 1-8. Paper 72.

[C19] Mráz Albert, Fazekas Péter, "Analysis of Channel Estimation Imperfections Within the 3GPP LTE Physical Layer." In: *Second International Conference on the Network of the Future: NOF'2011*. Paris, France, 2011.11.28-2011.11.30. (IEEE) IEEE

[C20] Törös István, Fazekas Péter, "An energy efficient cellular mobile network planning algorithm." In: *IEEE 73rd Vehicular Technology Conference: VTC2011-Spring*. Budapest, Hungary, 2011.05.15-2011.05.18. (IEEE, IEEE VTS)5 p. Paper 1.

[C21] Törös István, Fazekas Péter, "Planning and network management for energy efficiency in wireless systems." In: *Future Network and Mobile Summit 2011*. Warsaw, Poland, 2011.06.15-2011.06.17. 8 p. Paper 54.

## Technical reports

[T1] Fazekas Péter, Imre Sándor, Iván Gábor, Fülöp Attila, Zsíros Attila, "Első rész: HSDPA teljesítőképesség", *Mobiltelefon hálózatok csomagkapcsolt forgalom-modellézése és méretezése a hozzáférési és az átviteli hálózatrészekben, HSDPA hálózatok minőségi kérdései*, BME Department of Telecommunications - Tmobile bilateral research project, report, October, 2006

[T2] Fazekas Péter, Imre Sándor, Pap László, Iván Gábor, Fülöp Attila, Zsíros Attila, Németh Zoltán, "Második rész: Hálózat optimalizálási és méretezési kérdések", *Mobiltelefon hálózatok csomagkapcsolt forgalom-modellézése és méretezése a hozzáférési és az átviteli hálózatrészekben, HSDPA hálózatok minőségi kérdései*, BME Department of Telecommunications - Tmobile bilateral research project, report, October, 2006

[T3] Fazekas Péter, Jeney Gábor, "Második rész: 2G-3G adatforgalommal kapcsolatos és kapacitás-méretezési feladatok" *Mobiltelefon hálózatok csomagkapcsolt forgalom-modellézése és méretezése a hozzáférési és az átviteli hálózatrészekben, Rádióhálózat tervezés kutatási és fejlesztési feladatok*, BME Department of Telecommunications - Tmobile bilateral research project, report, October, 2006

# Appendix B

## New scientific results formulated in the dissertation

### Result group 1: Analytical modelling framework of broadband cellular networks with multi-rate traffic sources

#### Result 1.1

I have developed a general connection level stochastic modelling framework of a radio cell of broadband wireless networks. The model has the following novel capabilities:

- it incorporates arbitrary connection duration distributions,
- it describes user mobility with arbitrary dwell times
- it models the variability and burstiness of user generated traffic
- it describes immediate rejection and service rate reduction admission policies.

The channel holding time is modelled by a phase type distribution, which is proven to have descriptors according to

$$\underline{t}^{(N,k)} = \underline{d}^{(R,k)} \otimes \underline{l}^{(R,k)} \quad \mathbf{T}^{(N,k)} = \mathbf{D}^{(R,k)} \oplus \mathbf{L}^{(k)},$$

where  $\underline{d}^{(R,k)}$ ,  $\mathbf{D}^{(R,k)}$  are the descriptors of type  $k$  customers' residual dwell time and  $\underline{l}^{(R,k)}$ ,  $\mathbf{L}^{(k)}$  are the descriptors of the session duration distribution. The service process of the system is

proven to have phase type service time distribution with descriptors

$$\underline{s}^{(N,k)} = \underline{t}^{(N,k)} \otimes \underline{q}^{(k)} \quad \mathbf{S}^{(N,k)} = \mathbf{T}^{(N,k)} \oplus \mathbf{Q}^{(k)},$$

where  $\mathbf{Q}^{(k)}$  is the infinitesimal generator of the Markov chain that describes generated bitrate pattern of a type  $k$  session and the service process defines phase dependent capacity requirements for the connections.

### Result 1.2

I have introduced a general Markovian source model that is able to capture the variability of generated user bitrate. The novelty of the model is that it describes active traffic generation phases, containing actual transmitted bursts; there can be gaps between bursts and bursts may be transmitted by different bitrates. The model describes inactive phases between active transmission of a session. The model enables the assumption of arbitrary Phase Type distributed durations of traffic generation phases and burst lengths. This source model is described by an underlying Markov chain, characterised by its infinitesimal generator and initial probability vector

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_{\text{eff}} & \underline{A}_{\text{eff}}^0 \cdot \underline{i} \\ \underline{I}^0 \cdot \underline{a}_{\text{eff}} & \mathbf{I} \end{bmatrix}, \quad \underline{q} = [\underline{a}_{\text{eff}}, \underline{i}].$$

where  $\underline{i}$  and  $\mathbf{I}$  are descriptors of the inactive idle periods and

$$\mathbf{A}_{\text{eff}} = \mathbf{A} \oplus \mathbf{A}^*, \quad \underline{a}_{\text{eff}} = \underline{a} \otimes \underline{a}^*,$$

where  $\underline{a}$  and  $\mathbf{A}$  are the descriptors of the duration of active transmission periods and

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{B}_{(1)} & P_{12} \cdot \underline{B}_{(1)}^0 \cdot \underline{b}_{(2)} & \dots & P_{1R} \cdot \underline{B}_{(1)}^0 \cdot \underline{b}_{(R)} \\ P_{21} \cdot \underline{B}_{(2)}^0 \cdot \underline{b}_{(1)} & \mathbf{B}_{(2)} & \dots & P_{2R} \cdot \underline{B}_{(2)}^0 \cdot \underline{b}_{(R)} \\ & \vdots & & \\ P_{R1} \cdot \underline{B}_{(R)}^0 \cdot \underline{b}_{(1)} & \dots & & \mathbf{B}_{(R)} \end{bmatrix}$$

$$\underline{a}^* = [p_1 \cdot \underline{b}_{(1)} \dots p_R \cdot \underline{b}_{(R)}],$$

where  $\underline{b}_{(r)}$  and  $\mathbf{B}_{(r)}$  are the descriptors of the durations of rate  $r$  bursts,  $r \in 1 \dots R$  and  $P_{r_1 r_2}$  is the probability that a burst with rate  $r_1$  is followed by a burst with rate  $r_2$ . (Section 2.3)

### Result 1.3

I have identified the state space and transitions of the queueing system describing the general model of broadband cellular networks. I described the state transitions for immediate blocking and rate reduction admission policies. (Chapter 3)

### Result 1.4

I have shown that the local balance equations of the form

$$\lambda_N \alpha_k s_i p(\underline{n}^*) + \sum_{j=1, j \neq i}^P (n_j^* + 1) S_{ji} \cdot p(\underline{n}^* + \underline{e}_j) = (n_i^* + 1) \cdot p(\underline{n}^* + \underline{e}_i) \left( S_i^0 + \sum_{j=1, j \neq i}^P S_{ij} \right).$$

hold within the non-blocking part of the state space. Based on this, I have elaborated the modified version of the Kaufman-Roberts recursive algorithm for the calculation of the channel occupancy probability  $p(m)$ , which I gave as  $\tilde{p}(m) = 0$  for  $m < 0$ ,  $\tilde{p}(0) = 1$ , and for  $m > 0$

$$\tilde{p}(m) = \sum_{k=1}^K \sum_i \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{(k),N}(m - r_i^{(k)}) + \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{(k),H}(m - r_i^{(k)})$$

and

$$p(m) = \tilde{p}(m) \frac{1}{\sum_{m=0}^{C_0} \tilde{p}(m)}.$$

I used this algorithm to obtain approximate results of session level performance parameters of the modelled system. I compared the analytical results with simulations and found that the accuracy of the approximation is suitable. (Sections 3.4, 3.5)

## Result group 2: Residual lifetime of handover connections in cellular systems

### Result 2.1

I have developed a general method that is applicable to determine the distribution of the residual session duration of a customer that arrives to a given radio cell with handover. I have shown that for customers having dwell time distribution with density  $g(t)$  and residual dwell time

density  $g_R(t)$  and session duration distribution  $F(t)$ , the residual session duration after handover is given as

$$F_R(t) = \sum_{i=1}^I p^{(i)} \cdot F_R^{(i)}(t),$$

where

$$F_R^{(i)}(t) = \frac{\int_0^\infty (F(t+x) - F(x))g^{(i)}(x)dx}{\int_0^\infty (1 - F(x))g^{(i)}(x)dx},$$

and

$$g^{(i)}(x) = g_R(x) * g(x) * \dots * g(x).$$

and

$$p^{(i)} = \frac{N_{HO}^{(i)}}{N_{HO}} = \frac{\sum_{k:H(k,1) \leq i} B_k \cdot \Pr(\tau_S > \tau^{(i)}) \cdot \sum_{all \underline{r}^{(i)}(k)} Pr(\underline{r}^{(i)}(k))}{\sum_{j=1}^I \sum_{k:H(k,1) \leq j} B_k \cdot \Pr(\tau_S > \tau^{(j)}) \cdot \sum_{all \underline{r}^{(j)}(k)} Pr(\underline{r}^{(j)}(k))} = \frac{\hat{p}^{(i)}}{\sum_{j=1}^I \hat{p}^{(j)}}.$$

I have validated the method using computer simulations.(Sections 4.2, 4.4)

## Result 2.2

I have developed the method of determining the residual connection lifetime distribution when user describing time parameters have, or approximated by phase-type distributions. I have shown that in case of phase-type session duration, the residual distribution has the same phase structure, only the initial probability vector changes according to

$$\underline{l}_R^{(i)} = \frac{\int_0^\infty \underline{l} \cdot e^{Lx} \cdot g^{(i)}(x)dx}{\int_0^\infty \underline{l} \cdot e^{Lx} \cdot \underline{h} \cdot g^{(i)}(x)dx}.$$

Using computer simulations I have shown that the phase-type approximation introduces tolerable inaccuracy into the method.(Sections 4.2.2, 4.4)

## Result 2.3

I have investigated two specific scenarios, the motorway and the homogeneous cell structure scenarios and I have shown that the residual session length has the same distribution in these cases. (Section 4.3)

## Result group 3: Capacity and performance of 3G systems

### Result 3.1

I introduced a new interpretation of the average useful capacity of a UMTS radio cell as

$$\bar{R}_{\text{UMTS}} = N \cdot \sum_{k=1}^K n_k \cdot Rb_k,$$

where  $N$  is the average number of scheduled radio interface connections,  $n_k$  is the ratio of type  $k$  connections,  $Rb_k$  is the useful datarate of a type  $k$  radio bearer. I have developed the calculation method of this capacity, that is

$$N = \frac{P_{\text{avg}}^0 - P_{\text{Pil}}^0}{\sum_{k=1}^K n_k \cdot \bar{P}_k^0},$$

where  $P_{\text{avg}}^0$  is the average power used for data transmission,  $P_{\text{Pil}}^0$  is the power of pilot and control channels and  $\bar{P}_k^0$  is the expectation of the power used to serve a type  $k$  bearer. The latter should be calculated as

$$\bar{P}_k^0 = \varepsilon_k \cdot \frac{Rb_k}{Rc} \cdot \left( (1 - \bar{\rho}_k) \cdot P_{\text{avg}}^0 + \sum_{b \neq 0} P_{\text{avg}}^b \cdot \bar{f}_k^b + \bar{\eta}_k \right),$$

where  $\varepsilon_k$  is the required SINR (Signal to Interference Noise Ratio) level of bearer  $k$ ,  $Rc$  is the chiprate of the system,  $\bar{\rho}_k$  is the average orthogonality factor,  $P_{\text{avg}}^0$  and  $P_{\text{avg}}^b$  are the average power levels of the serving and  $b$ th neighbouring base stations,  $\bar{f}_k^b$  and  $\bar{\eta}_k$  are the average of the path loss ratios and relative noise power.

I have shown by means of comparison to the results of snapshot simulations, that the calculation method is applicable with reasonable accuracy. (Sections 5.2.2, 5.3.1)

### Result 3.2

I have introduced a method that is capable of determining the average used power level of a UMTS Node B, if the average amount of carried traffic is given. The method is composed of the following steps:

$$N = \frac{\bar{R}_{\text{UMTS}}}{\sum_{k=1}^K n_k \cdot Rb_k},$$

where  $\bar{R}_{\text{UMTS}}$  is the given amount of average carried traffic. Then the linear system

$$\bar{P}_k^0 = \varepsilon_k \cdot \frac{Rb_k}{Rc} \cdot \left( (1 - \bar{\rho}_k) \cdot \left( N \cdot \sum_{l=1}^K n_l \bar{P}_l^0 + P_{\text{Pil}}^0 \right) + \sum_{b \neq 0} P_{\text{avg}}^b \cdot \bar{f}_k^b + \bar{\eta}_k \right). \quad (\text{B.-18})$$

is solved for all  $\bar{P}_k$ . The total average used power is then

$$P_{\text{avg}}^0 = N \cdot \sum_{k=1}^K n_k \cdot \bar{P}_k^0 + P_{\text{Pil}}^0. \quad (\text{B.-18})$$

I proposed an iterative method to obtain the average used power level, if the neighbouring (interfering) base stations are modelled as carrying the same amount of traffic. The method consists of the following steps

- Step 0. Suppose arbitrary level of interfering Node B powers (less than the maximal output power).
- Step 1. With the given interfering power level solve (B). Determine the used power of the Node B in question with (B).
- Step 2. Substitute the resultant power level as interfering power. Repeat step 1 and step 2 until convergence.

(Section 5.2.2)

### Result 3.3

I developed a method to analyse HSDPA useful throughput and average cell throughput of HSDPA services. The method is the following. SINR is determined according to

$$\text{SIR}_{\text{HS}}(r, \phi) = 16 \cdot \frac{P_{\text{HS}}^0 \cdot L_i^0(r, \phi)}{(1 - \rho(r)) \cdot P_{\text{inst}}^0 \cdot L_i^0(r, \phi) + \sum_{b \neq 0} P_{\text{inst}}^b \cdot L_i^b(r, \phi) + P_{\text{noise}}},$$

where  $P_{\text{HS}}^0$  is the power allocated for HSDPA transmission (considered as remaining power left unused by Release'99 services),  $P_{\text{noise}}$  is the power of the thermal noise, other quantities are as in previous equations. It is assumed that SINR is mapped to achievable useful datarate denoted by  $R_i(r, \phi)$ . The average throughput achievable by a category  $i$  HSDPA terminal is then calculated as:

$$R_i = \int_{r=0}^R \int_{\phi=0}^{2\pi} R_i(r, \phi) g_{\text{HS}}(r, \phi) d\phi dr,$$

where  $g_{\text{HS}}(r, \phi)$  is the probability density function of the HSDPA user spatial distribution over the plain. The average HSDPA capacity is

$$R_{\text{HSDPA}} = \sum_{i=1}^{12} R_i \cdot \alpha_i,$$

where  $\alpha_i$  is the ratio of category  $i$  HSDPA devices among all devices. (Sections 5.2.3, 5.3.2)

# Appendix C

## Theorems and proofs

### C.1 Proof of the existence of local balance equations (3.7)

**Theorem** The local balance equations (3.7) hold in the non-blocking part of the state space (and in the whole state space of an infinite capacity system).

**Proof.** The global balance equation equating the incoming and leaving rates of state  $\underline{n}^*$  has the form of

$$p(\underline{n}^*) \left( \lambda_N \alpha_k - \sum_{i=1}^P n_i^* S_{ii} \right) = \sum_{i=1}^P \lambda_N \alpha_k s_i p(\underline{n}^* - \underline{e}_i) + \sum_{i=1}^P (n_i^* + 1) S_i^0 p(\underline{n}^* + \underline{e}_i) + \sum_{j=1}^P (n_j^* + 1) \sum_{i=1, i \neq j}^P S_{ij} p(\underline{n}^* + \underline{e}_i - \underline{e}_j) \quad (\text{C.1})$$

An equivalent local balance equation with (3.7), expressing the balance around state  $\underline{n}^*$  is

$$\lambda_N \alpha_k s_i p(\underline{n}^* - \underline{e}_i) + \sum_{j=1, j \neq i}^P (n_j^* + 1) S_{ji} p(\underline{n}^* + \underline{e}_j - \underline{e}_i) = n_i^* p(\underline{n}^*) \left( S_i^0 + \sum_{j=1, j \neq i}^P S_{ij} \right). \quad (\text{C.2})$$

Due to the properties of Markov chains  $S_i^0 + \sum_{j=1, j \neq i}^P S_{ij} = -S_{ii}$  in the right hand side of (C.2), thus  $-n_i^* S_{ii} p(\underline{n}^*)$  can be substituted into (C.1). After substitution and cancelling the

obviously equal terms, we arrive to

$$\lambda_N \alpha_k p(\underline{n}^*) + \sum_{i=1}^P \sum_{j=1, j \neq i}^P (n_j^* + 1) S_{ji} p(\underline{n}^* + \underline{e}_j - \underline{e}_i) = \quad (C.3)$$

$$\sum_{i=1}^P (n_i^* + 1) S_i^0 p(\underline{n}^* + \underline{e}_i) + \sum_{j=1}^P (n_j^* + 1) \sum_{i=1, i \neq j}^P S_{ij} p(\underline{n}^* + \underline{e}_i - \underline{e}_j).$$

The second terms of both sides of this equation are equal, i.e.

$$\sum_{i=1}^P \sum_{j=1, j \neq i}^P (n_j^* + 1) S_{ji} p(\underline{n}^* + \underline{e}_j - \underline{e}_i) = \sum_{j=1}^P (n_j^* + 1) \sum_{i=1, i \neq j}^P S_{ij} p(\underline{n}^* + \underline{e}_i - \underline{e}_j), \quad (C.4)$$

since both sides sum up the same quantities but in different orders. Cancelling these terms (C.3) gets the form

$$\lambda_N \alpha_k p(\underline{n}^*) = \sum_{i=1}^P (n_i^* + 1) S_i^0 p(\underline{n}^* + \underline{e}_i). \quad (C.5)$$

Substituting  $(n_i^* + 1) S_i^0 p(\underline{n}^* + \underline{e}_i)$  from (3.7), using that  $\sum_{i=1}^P s_i = 1$  and applying the same argument that was used regarding (B.3), we get

$$\lambda_N \alpha_k p(\underline{n}^*) = \lambda_N \alpha_k p(\underline{n}^*),$$

hence the local balance equations of the form (3.7) holds.  $\square$

## C.2 Alternative proof of the existence of equilibrium distribution 3.12

As it was stated in Section 3.3 that the queueing system model described in analogue with a BCMP network containing infinite server (or type 3) queues, in terms of state transitions and rates in the non-blocking subspace; considering unbounded system capacity the system is equivalent with a BCMP network of IS queues. It was derived that the system has the equilibrium probability distribution of (3.12), where  $F_i$ -s are elements of  $\underline{F}$  from (3.10).

This can be shown from the BCMP point of view, as follows. As in Section 3.3 the states and transitions considering a class  $k$  new connection were described, here we refer to this case as well. This means that the incoming rate of the corresponding BCMP network is  $\lambda_N \cdot \alpha_k$ . We

know that the joint queue length distribution of the BCMP network with exponential IS queues with service rates of  $\mu_i$  has the form of:

$$p(\underline{n}^*) = \frac{1}{G} \prod_{i=1}^P \left( \frac{\lambda_i}{\mu_i} \right)^{n_i^*} \frac{1}{n_i^*!}, \quad (\text{C.6})$$

where the individual arrival rates  $\lambda_i$  are calculated from the traffic equations, namely

$$\lambda_i = \lambda_N \cdot \alpha_k \cdot \pi_{01} + \sum_{j=1, j \neq i}^P \lambda_j \cdot \pi_{ji}. \quad (\text{C.7})$$

Here  $\pi_{0i}$  denote the probability that an arrival jumps into queue  $i$  and  $\pi_{ji}$  is the probability that a job enters queue  $i$  after finishing in queue  $j$ . As each queue of this analogous BCMP system represents a phase of the service time of the system considered (with initial probability vector  $\underline{s}$  and rate matrix  $\mathbf{S}$ ), the parameters in (B.7) are:  $\pi_{0i} = s_i$ ,  $\pi_{ji} = \frac{S_{ji}}{-S_{jj}}$  and in (B.6)  $\mu_i = -S_{ii}$ .

Collecting all  $\lambda$ -s in (B.7) into one side and writing the equation into vectorial form, we get:

$$-\lambda_N \cdot \alpha_k \cdot \underline{s} = \underline{\lambda} \cdot \text{diag}\left(\frac{1}{-S_{jj}}\right) \cdot \mathbf{S}, \quad (\text{C.8})$$

where  $\underline{\lambda}$  contains  $\lambda_i$ -s and  $\text{diag}\left(\frac{1}{-S_{jj}}\right)$  is a diagonal matrix containing  $\frac{1}{-S_{jj}}$  values. From this we have

$$\underline{\lambda} = -\lambda_N \cdot \alpha_k \cdot \underline{s} \cdot \mathbf{S}^{-1} \cdot \text{diag}(-S_{jj}). \quad (\text{C.9})$$

Comparing this with (3.10), we see that

$$\underline{\lambda} = \underline{F} \cdot \text{diag}(-S_{jj}). \quad (\text{C.10})$$

If we substitute each  $\lambda_i$  into (B.6), we get back to the original statement (3.12).

# Appendix D

## List of Abbreviations

- 3GPP - Third Generation Partnership Project
- ARQ - Automatic Repeat reQuest
- ATM - Asynchronous Transfer Mode
- CCPCH - Common Control Physical Channel (UMTS radio interface)
- CDMA - Code Division Multiple Access
- CIR - Carrier to Interference Ratio
- CPICH - Common Pilot Channel (UMTS radio interface)
- CQI - Channel Quality Indicator
- CTMC - Continuous Time Markov Chain
- FDD - Frequency Division Duplex
- FTP - File Transfer Protokol
- GPRS - General Packet Radio Service
- GSM - General System for Mobile communications
- HSCSD - High Speed Circuit Switched Data
- HSDPA - High Speed Downlink Packet Access
- HS-DSCH - High Speed Downlink Shared Channel
- HSUPA - High Speed Uplink Packet Access
- IP - Internet Protocol
- IS - Infinite Server
- LTE - Long Term Evolution, the next generation radio interface of 3GPP
- MMPP - Markov Modulated Poisson Process
- Node B - base station in 3GPP terminology

OF - Orthogonality Factor  
OFDMA - Orthogonal Frequency Division Multiple Access  
OVSF - Orthogonal Variable Spreading Factor codes  
QAM - Quadrature Amplitude Modulation  
QoS - Quality of Service  
Release'99, R'99 - 3G network without extensions defined in later standard releases  
RNC - Radio Network Controller  
SIR - Signal to Interference Ratio  
SMTP - Simple Mail Transfer Protocol  
SNR - Signal to Noise Ratio  
TCP - Transmission Control Protocol  
TDMA - Time Division Multiple Access  
UMTS - Universal Mobile Telecommunication System  
UTRA - UMTS Terrestrial Radio Access  
UTRAN - UMTS Terrestrial Radio Access Network  
WATM - Wireless Asynchronous Transfer Mode  
WCDMA - Wideband Code Division Multiple Access